Published in final edited form as:

Nat Protoc. ; 6(12): 1969–1980. doi:10.1038/nprot.2011.417.

Identifying microbial fitness determinants by Insertion Sequencing (INSeq) using genome-wide transposon mutant libraries

Andrew L. Goodman^{1,2}, Meng Wu¹, and Jeffrey I. Gordon¹

¹Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108

Abstract

Insertion Sequencing (INSeq) is a method for determining the insertion site and relative abundance of large numbers of transposon mutants in a mixed population of isogenic mutants of a sequenced microbial species. INSeq is based on a modified mariner transposon containing MmeI sites at its ends, allowing cleavage at chromosomal sites 16–17bp from the inserted transposon. Genomic regions adjacent to the transposons are amplified by linear PCR with a biotinylated primer. Products are bound to magnetic beads, digested with MmeI, and barcoded with sample-specific linkers appended to each restriction fragment. After limited PCR amplification, fragments are sequenced using a high-throughput instrument. The sequence of each read can be used to map the location of a transposon in the genome. Read count measures the relative abundance of that mutant in the population. Solid-phase library preparation makes this protocol rapid (18h), easy to scale-up, amenable to automation, and useful for a variety of samples. A protocol for characterizing libraries of transposon mutant strains clonally arrayed in multi-well format is provided.

Keywords

bacteria; modified mariner transposon; solid-phase library preparation; robotics; massively parallel DNA sequencing; combinatorial pooling and mapping of clonally arrayed transposon mutant libraries

INTRODUCTION

Development of the protocol

By exposing populations of transposon mutants to selective conditions *in vitro* or *in vivo*, it is possible to identify changes in the representation of mutants, thereby allowing identification of genes and pathways that are key to fitness under the conditions being examined. In principle, this approach could be applied to any microbe whose genome has been sequenced and that is amenable to basic genetic manipulation (insertion of foreign

Address correspondence to andrew.goodman@yale.edu or jgordon@wustl.edu.

²Current address: Section of Microbial Pathogenesis and Microbial Diversity Institute, Yale School of Medicine, New Haven CT 06536

AUTHOR CONTRIBUTIONS

A.L.G. and M.W. designed the experiments and software. A.L.G., M.W. and J.I.G. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing financial interests.

DNA, antibiotic selection). Insertion Sequencing (INSeq) combines existing methods for signature-tagged mutagenesis¹ with new techniques for high-throughput DNA sequencing such that the precise location and relative abundance of tens of thousands of mutants can be determined in parallel. This approach is based on modification of the broad host-range, randomly integrating, transposable element, *mariner*, to allow short fragments of genomic DNA adjacent to the site of transposon insertion to be captured, sequenced, and quantified². Mutants that decrease in relative abundance in a selective condition likely are in genes important for fitness under that condition; mutants that increase in abundance under the selection highlight genes that could be deleterious. The protocol described here includes significant improvements from our earlier publication²: because library preparation can now be performed in 96-well, bead-based format, reagent costs and sample losses are minimized, and large numbers of samples can be barcoded and sequenced in a single sequencing run.

Applications of the method

Because INSeq library preparation can be completed in multiplex format, experiments can involve material obtained from high resolution time-series studies of the impact of *in vitro* or *in vivo* selections, and from multiple biological and technical replicates. Any microbe whose genome has been sequenced and that is amenable to *mariner* transposon mutagenesis could potentially be investigated with INSeq. This method can also be used to efficiently determine the insertion site of transposon mutant strains that have been clonally arrayed in multi-well format (see Box 1). This latter technique relies on a combinatorial pooling approach: a small number of pools are defined, and a liquid handling robot is used to place each mutant strain into a subset of these pools in a unique pattern. The pools are then characterized by INSeq, and the presence/absence pattern of a given insertion site across the pools reflects the unique pattern associated with the corresponding bacterial strain in the multi-well collection. Alternative pooling patterns that use larger numbers of pools to further reduce the likelihood of mistaking one strain for another can also be considered^{3–5}. These alternatives would be appropriate if many mutants in the population were likely to carry insertions at the same location in the genome, but require greater effort in preparing libraries because of the increased number of pools. Furthermore, these alternate techniques were not developed specifically for mapping transposon insertions and may require modification in order to accomplish this goal.

Box 1

Additional protocol: combinatorial pooling and mapping of clonally arrayed transposon mutant libraries by INSeq

TIMING: Pooling strains on an EpMotion liquid handling robot takes ~1.6h/96-well tray. Additional time is required for preparing the arrayed mutant collection by manually or robotically picking transposon-mutagenized colonies into 96-well trays and conducting INSeq library preparation on the pooled strains.

1. Prepare EpMotion programs. Two sets of pooling patterns are provided in the analysis package INSeq_analysis/Arrayed_library directory (Supplementary Dataset 1): “16384_strings.txt” is the list of 24-bit strings with a minimum Hamming distance of 6. “13000_strings.txt” is the subset of these that are less likely to be mistaken for each other if multiple archived strains happen to carry transposons at the identical genome coordinate. The larger set can be used to map up to 170 96-well trays, while the smaller set is preferable for arrayed libraries of under 135 trays. A script called write_dws.pl translates these pooling patterns into programs for the EpMotion robot. Each program will pool 5 trays (20 programs are needed for a 100-tray mutant library, for example).

To create EpMotion .dws files, go to the data analysis package INSeq_analysis/Arrayed_library directory (Supplementary Dataset 1) and type:

```
perl write_dws.pl <pooling pattern> <number of programs>
```

This will create an output directory called “Write_dws_output_X_programs”. Create a new application on the EpMotion, and then import the .dws files in this directory into the new application and confirm that they are functioning properly. (Note that the provided script is designed for Midwest Scientific Cat. No. TP92696 96-well culture trays and the EpMotion Thermorack for 24 cryotubes, Cat. No. 960002491. Alternatives may require changes to the program). The write_dws.pl script will also create a map file that will be used in the data analysis step. Although the software provided is designed for the EpMotion instrument, analogous programs could be written for other liquid-handling robots if they can be run from a text-format file.

2. Prepare mutant collection by manually or robotically picking transposon-mutagenized colonies into 96-well trays containing 250µL culture media and appropriate antibiotics. Incubate under standard conditions to allow cultures to grow to turbidity, and note any wells with poor or no growth.
3. For each 96-well culture tray, prepare two 96-well archive trays by adding 30µL of sterile culture medium plus 40% glycerol to each well. Transfer 30µL of the turbid culture to the corresponding well in each of these archive trays, mix well, and seal with foil lid before storing archive trays at −80°C. Turbid cultures may need to be mixed before dispensing into archive trays. To recover strains from the archive trays, wipe down the foil lid with 70% ethanol and puncture the well of interest with a sterile pipette tip. Draw a small amount of frozen material onto the tip of the pipette without allowing the entire plate to thaw. Re-streak this material onto agar plates, and place a small foil patch over the targeted well in the archive tray before returning to the −80°C freezer.
4. Seal culture trays in groups of five with Parafilm and store at 4°C.
5. On the EpMotion, initiate the Pool_group_1.dws program. Set up the robot deck as shown in the graphical interface, with tips, input culture trays #1–5, and a cryovial rack containing 24 2mL screw-cap cryovials (lids removed and vials labeled with program and pool number). In the cryovial rack, tubes in the top row should be labeled 1–6, 2nd row 7–12, etc. While the program is running, monitor the waste tip bucket to ensure that it doesn’t overflow. When the program is complete, cap the 2mL output cryovials and store at −20°C. **CRITICAL STEP:** each program takes ~8h to complete. Consider adding a growth inhibitor if your mutants are likely to grow during this process. If the robot is performing without interventions required, programs can be run overnight. Also, if input trays have been stored at 4°C, cultures may need to be resuspended manually before being placed in the robot. See Supplementary Movie 2 for a demonstration of the pooling procedure.
6. Run pooling programs on each subsequent set of 5 culture trays as described in Box 1, Step 5.
7. When pooling is complete, thaw all cryovials, vortex well, and combine half of the volume of each cryovial with same number from each pooling program to create 24 final pools. Regardless of the number of programs run, there should be 24 pools at this point. Save the remainder of each cryovial at −20°C. **CRITICAL STEP:** be sure to combine like pools and NOT to combine different pools from a single program.

8. Prepare INSeq libraries from each of the 24 pools (Steps 1–78), associating a different barcode with each pool.
9. Run data analysis pipeline as described in the arrayed library section of the README.txt included in the analysis package.

CRITICAL STEP: the cutoffs for presence/absence of a strain in a pool may need to be determined empirically by examining the distributions of reads across pools for a given dataset. As a starting point, normalize counts/pool for each strain by the overall number of reads in the pool; for each strain, set the presence/absence of a pool to 0 if there are less than 3 reads associated with that pool and to 1 if there are 3 or more reads associated with the pool.

Limitations and alternative techniques

In bacterial genomes, a high percentage (>90%) of insertions can be precisely localized using INSeq. With more complex organisms, the short (16–17bp) fragments of adjacent chromosomal sequence captured and quantified by MmeI digestion may be insufficient to uniquely identify transposon locations. As with any negative selection technique, stochastic bottlenecks in the selective process need to be considered when determining the optimal number of mutants to be included in an experiment. Because each insertion is mapped with nucleotide-level resolution, comparing the behavior of multiple mutants in the same gene highlights these selective bottlenecks.

Other techniques for mapping and quantifying transposon insertions by high-throughput sequencing are also available^{6–10}. They include two protocols named Tn-Seq: one version^{8,10} also uses an MmeI-adapted *mariner* transposon but does not include a transposon-specific, linear PCR step during library preparation. Compared to INSeq, this simpler approach to library preparation could reduce the effort required for sequencing. The other Tn-Seq protocol⁹ does not use an MmeI-adapted *mariner* transposon; instead, it employs a circularization step to capture the targeted DNA fragments. This could allow use of transposons that cannot be modified to include MmeI sites in their inverted repeats. Additional approaches, termed “High-throughput Insertion Tracking by deep Sequencing (HITS)”⁶ or “Transposon-Directed Insertion-Site sequencing (TraDIS)”⁷ also do not require MmeI sites in the transposon. For these protocols, mutagenized DNA is randomly sheared, adapters are ligated to all fragments, and transposon-chromosomal junctions are enriched by PCR either with (HITS) or without (TraDIS) an affinity-purification step. In principle, these MmeI-independent methods could allow capture of longer fragments of genomic sequence adjacent to each transposon, a feature which may be necessary for mapping insertions in more complex genomes.

The INSeq protocol we describe below has several advantages compared to these other methods: (1) introducing an MmeI site in the transposon allows the excised transposon/chromosomal junctions to be of uniform size, thus avoiding size preferences during PCR; (2) the linear PCR step at the beginning of the protocol enriches the target and increases the efficiency of library preparation; (3) the solid phase-based technique separates target sequences from chromosomal DNA background, making it easy to scale up to 96-well plate format; (4) robotic automation makes the protocol amenable to monitoring changes in the representation of mutants during studies that require repeated, high resolution sampling; and (5) descriptions for modifying the INSeq protocol to map arrayed mutant libraries are provided in Box 1. If other transposons that cannot be modified with MmeI sites are used, alternative protocols such as the Tn-seq method described by Gallagher *et al.*⁹, HITS⁶, or TraDIS⁷ are preferable.

Experimental Design

Overview—An overview of the INSEQ protocol is provided in Figure 1. Steps 1–16 describe the procedure used for extraction of DNA. In Steps 17–20, DNA is further purified using columns. Steps 21–23 describe a linear PCR procedure that is designed to enrich the transposon/chromosomal junction regions by using a primer complementary to the transposon-specific sequence. Steps 21–23 are necessary for two reasons. First, they add a biotin tag to each target molecule which facilitates high-throughput, low-volume solid-phase library preparation. Second, the linear PCR enriches for the desired target molecules early in the procedure, minimizing the amount and cost of enzymes needed for subsequent reaction steps. In Steps 24–36, the linear PCR product is bound to magnetic beads, allowing the PCR product to be separated from ‘background’ chromosomal DNA by affinity capture. Second strand synthesis (and creation of the double-stranded MmeI site) takes place in Steps 37–43; because MmeI requires two sites for efficient DNA cleavage,¹⁴ a second MmeI recognition site is provided in trans via the dsDNA fragment M12. Steps 44–52 use MmeI to cut the double-strand DNA fragment to a uniform length (64–65 bp). The product of this reaction consists of the amplified transposon region generated by the linear PCR plus 16–17bp of chromosomal DNA. In Steps 53–63, sequencing adapters are ligated to the transposon/chromosomal junction to introduce the complementary sequence for the sequencing primer employed by the Illumina Genome Analyzer/HiSeq instrument. By varying the four-nucleotide barcode sequences in the adapter, multiple samples can be sequenced in a single lane of an Illumina flow cell. With limited cycles of PCR (Steps 64–78), targets are enriched in the linear range. PCR products are then purified by gel-extraction to remove extra primer and adapters and sequenced. Data analysis is described in Steps 79–86; detailed documentation of data analysis tools is available in the README.txt in the INSeq_analysis package (see Supplementary Data 1). Box 1 describes an additional protocol for combinatorial pooling and mapping of clonally arrayed transposon mutant libraries by INSeq.

Starting material—Steps 1–16 describe the procedure used for extraction of DNA from the transposon library (the ‘input mutant population’) and from biological specimens containing mutants that survive a given selection (the ‘output mutant population’). These steps for DNA isolation were originally developed for mouse feces and for samples obtained from more proximal regions of the murine gastrointestinal tract such as the cecum, and can also be used for *in vitro* mutant populations. Transposon mutant libraries can be introduced into a wide variety of biological systems (e.g. various animal or environmental habitats) or characterized in many different *in vitro* formats (e.g., chemostats; tissue culture). In these cases, users should isolate crude DNA using protocols designed for the specific system chosen for study. Whatever the system, experiments need be designed so that the amount of input and output material is sufficient for library preparation. For example, if the mutagenized species constitutes 100% of a microbial community, then the initial purification steps as described in the Procedure ideally should yield >500ng of DNA. If the mutagenized species represents only 50% of the population, then amount of starting material should be doubled.

Oligonucleotide sequence design—The linear PCR primer BioSamA is 5’-biotinylated and encodes a TEG spacer to reduce steric hindrance during the solid-phase enzymatic steps. The primer also encodes a transposon-specific region with a 26bp sequence complementary to sequences found on both sides of the transposon (underlined in Supplementary Table 1). Adaptors encode 4-basepair barcodes (marked in bold in Supplementary Table 1) that allow sample multiplexing and produce a 2bp NN overhang (complementary to the overhang produced by MmeI digestion) for ligation. For each adaptor, two complementary oligonucleotide sequences are provided in Supplementary

Table 1: LIB_AdaptT_x and LIB_AdaptB_x, where x is the unique barcode specified by the adaptor). These pairs of oligonucleotide sequences are annealed to each other during the Procedure. The LIB-PCR primers encode the necessary sequences for Illumina library preparation and sequencing.

Replicates and controls—Biological and technical replicates should be performed. The number of biological replicates required depends on the experimental system under study and should be determined empirically. Technical replicates can include identical samples associated with different sample-specific barcodes. The number of technical replicates should be determined by the user; two per biological sample is likely sufficient. A ‘no-DNA’ negative control will help identify cross-contamination between samples; these controls should be run in parallel with DNA-containing samples. Cross-contamination can occur if beads transfer from tube lids to the users’ gloves to other samples and can be avoided with careful technique.

MATERIALS

Reagents

Phenol:Chloroform:Isoamyl Alcohol 25:24:1, pH 7.9 (Applied Biosystems, Cat. No. AM9712)

- Caution: hazardous. Handle properly during experiments.

3M sodium acetate, pH 5.5 (Applied Biosystems, Cat. No. AM9740)

100% Isopropanol (Fisher Scientific, Cat. No. A416-4)

- Caution: flammable. Handle properly during experiments.

100% Ethanol (pharmco-AAPER, Cat. No. 111ACS200)

- Caution: flammable. Handle properly during experiments.

TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7; Ambion, Cat. No. AM9861)

NaCl (NaCl powder, Fisher Scientific, Cat. No. S271-10)

Tris (Trizma® base, Sigma-Aldrich, Cat.No. T1503-1KG)

EDTA (EDTA powder, Sigma-Aldrich, Cat.No. s657-500)

RNase A (Qiagen, Cat. No. 19101)

QIAquick PCR Purification Kit (Qiagen, Cat. No. 28106)

Platinum Pfx (Invitrogen, Cat. No. 11708-021)

10mM dNTPs (Invitrogen, Cat. No. 18427-013)

Dynabeads M-280 Streptavidin (Invitrogen, Cat. No. 112-05D)

Hexanucleotide Mix (Roche, Cat. No. 11277081001)

Klenow Fragment (3' - 5' exo-; 5 U/μL) (New England Biolabs, Cat. No. M0212S)

MmeI (New England Biolabs, Cat. No. R0637S)

T4 DNA ligase (2,000,000 U/mL) (New England Biolabs, Cat. No. M0202T)

Agarose (Roche, Cat. No. 11388991-001)

Electrophoresis buffer (TBE, Borate Buffer¹¹, or equivalent)

250mM NaOH (Fisher Scientific, Cat. No. S318 3)
 Boric Acid (Sigma-Aldrich, Cat. No. B7901-500)
 10bp DNA Ladder (Invitrogen, Cat. No. 10821-015)
 GelGreen Nucleic Acid Stain, 10,000x stock prepared in water (Biotium, Cat. No. 41005)
 20% Ficoll 400 (Sigma-Aldrich, Cat. No. F4375)
 0.1M disodium EDTA (Sigma-Aldrich, Cat. No. BP120-500)
 0.25% xylene cyanol (Sigma-Aldrich, Cat. No. X4126-10G)
 0.25% bromophenol blue; (Sigma-Aldrich, Cat. No. 114391-5G)
 QIAquick Gel Extractio Kit (Qiagen, Cat. No. 28706)

Primer sequences are provided in Supplementary Table 1. Primers were obtained from Integrated DNA Technologies (IDT).

For arrayed library mapping protocol in Box 1 only:

Glycerol (J.T. Baker, Cat. No. 2136-01)

Equipment

Zirconium beads, 0.1mm diameter (BioSpec Products, Cat. No. 11079101z)
 Screw-top cryovials, 2mL (Axygen Products, Cat. No. SCT-200-SS-C-S)
 Phase Lock Gel Light-2mL (5 Prime, Cat. No. 2302820)
 BeadBeater homogenizer (BioSpec Products)
 Refrigerated microcentrifuge
 Magnetic particle concentrator (MPC) for 1.5mL tubes (Invitrogen, Cat. No. 123-21D)
 MPC for PCR tubes (Invitrogen, Cat. No. 123-31D)
 Electronic 200µL 8-channel pipettor with variable speed control
 Heated oven or water bath set to 50°C
 SpeedVac or similar vacuum evaporator
 Thermocycler capable of 0.1°C/second ramp rate
 Electrophoresis equipment for agarose gels
 Non-UV based gel illuminator
 Invitrogen QuBit DNA spectrophotometer (or similar, for DNA concentrations ≤ 1 ng/
 uL)
 DNA sequencing facility with Illumina Genome Analyzer/HiSeq instrument

For arrayed library mapping protocol in Box 1 only:

EpMotion 5075PC liquid handling robot with TS50 single-channel dispensing tool and rack for 24 cryotubes

CRITICAL: Eppendorf, Cat. No. 960002491 was used, but temperature control is not necessary.

Note that with the EpMotion, each set of five plates in the arrayed library takes ~8h to pool.

Higher-throughput liquid handling robots would likely be faster.

EpMotion 50 μ L tips, nonfiltered, 1x96 tips per culture tray (Eppendorf, Cat. No. 960050200)

96-well culture trays (Midwest Scientific, Cat. No. TP92696)

The following data analysis resources are needed:

- .fasta format sequence file and .ptt format annotation file(s) for the genome of interest
- Bowtie short read aligner (Release 0.12.7 (ref. 12))
- INSeq data analysis package (INSeq_analysis; Supplementary Dataset 1)
- INSeq sample data package (INSeq_demo, downloadable from <http://gordonlab.wustl.edu/SuppData.html>)

CRITICAL: This package can be run on a computational cluster or a desk-top/laptop computer under a UNIX system with 4GB memory, perl and Bowtie installed. Basic understanding of Unix and perl are required for data analysis.

- (optional) Tab-delimited operon file in format

```
Gene A   PA,B
Gene B   PB,C
Gene C   PC,D
```

Where $P_{A,B}$ is a value between 0 and 1 (inclusive) reflecting the probability that Gene A and Gene B belong to the same operon¹³.

Reagent Setup (all buffers can be stored for months at room temperature)

2x buffer A. 200mM NaCl, 200mM Tris, 20mM EDTA. Adjust to pH 8 with concentrated HCl

2x Bind and Wash (B&W) buffer. 2M NaCl, 10mM Tris, 1mM EDTA. Adjust to pH 7.5 with concentrated HCl

1x Bind and Wash (B&W) buffer. Dilute 2X B&W buffer 1:1 with dH₂O

LoTE buffer. 3mM Tris, 0.2 mM EDTA. Adjust to pH 7.5 with concentrated HCl

Borate Buffer. 250mM NaOH. Adjust to pH 8-8.5 with Boric Acid

10x xylene cyanol DNA dye. 20% Ficoll 400, 0.1M disodium EDTA, 0.25% xylene cyanol

10x bromophenol blue DNA dye. 20% Ficoll 400, 0.1M disodium EDTA, 0.25% bromophenol blue

PROCEDURE

Isolation of Crude DNA TIMING: 4h

- 1 Add 250 μ L of zirconium beads to sample (e.g., ~10⁹ CFU of an *in vitro* mutant population or <600mg of cecal/fecal contents from mice colonized with a mutant population) in a 2mL screw-top vial (cryovial). These beads help with cell disruption.

- 2 Add 500µl 2x Buffer A, 210µl 20% SDS, and 500µl phenol:chloroform:isoamyl alcohol. Chill samples on ice.
- 3 Bead-beat on the “homogenize” setting for 2 min. Rest samples on ice for 2 min and repeat bead-beating for 2 min.
- 4 Centrifuge tubes in refrigerated microcentrifuge (4°C, 6,800 x g, 3 min).
- 5 Transfer aqueous phase (~600µl) to a Phase-Lock gel tube (these tubes should be pre-spun as described in the manufacturer’s instructions).
- 6 Add an equal amount (600µl) of phenol:chloroform:isoamyl alcohol to each Phase-Lock gel tube and mix by inversion.
- 7 Centrifuge tubes in a microcentrifuge (room temperature, 18,000 x g, 5 min).
- 8 Transfer aqueous phase to a microcentrifuge tube and discard organic phase.
- 9 Add 600µl of cold (–20°C) 100% isopropanol to microcentrifuge tube.
- 10 Add 60µl (1/10 volume) 3M sodium acetate (pH 5.5), and mix thoroughly by vortexing.
- 11 Incubate at –20°C for at least 1h. PAUSE POINT: samples can be stored overnight or longer at –20°C at this point.
- 12 Centrifuge tubes in a refrigerated microcentrifuge (4°C, 18,000 x g, 20 min). Carefully decant and discard supernatant.
- 13 Wash pellet with 500µl of 100% ethanol and centrifuge in a refrigerated microcentrifuge (4°C, 18,000 x g, 3 min). Carefully decant and discard supernatant.
- 14 Remove any excess ethanol by gently tapping the tube upside-down on a laboratory tissue.
- 15 Evaporate remaining supernatant in a vacuum evaporator (no heat, check at 5 min intervals until dry).
- 16 Resuspend pellet in 200µL TE buffer (pH 7). Incubate in oven or water bath at 50°C for 30 min, vortexing every ~10 min. The pellet will dissolve much faster once it has been dislodged from the wall of the microcentrifuge tube.

RNase treatment and clean-up of crude DNA using QIAquick columns TIMING: 1h

- 17 Transfer 100µL of the crude DNA sample to a new tube. Remainder should be stored at –20°C or –80°C.
- 18 Add 0.5µL RNase A and incubate at room temperature for 2 min. RNase treatment ensures accurate DNA quantitation in subsequent steps.
- 19 Clean up samples using QIAquick PCR purification columns following the manufacturer’s instructions. Make sure that all the ethanol-containing wash buffer is removed from the column before elution by pipetting around the inner rim with a 10µL tip, if necessary. Elute in 52µl of Buffer EB.
- 20 Measure the DNA concentration of each sample using a spectrophotometer (either a UV-based or dye-based system is acceptable). PAUSE POINT: samples can be stored for months at –20°C or –80°C at this point.

Linear PCR TIMING: 2h

- 21** Assemble the linear PCR reactions on ice:

Component	Volume/sample
dH ₂ O	to 100µL
Pfx Buffer	10µL
10mM dNTPs	2µL
50mM MgCl ₂	2µL
BioSamA (1 pmol/µL)	5µL
Clean DNA (From Step 20)	0.5-2µg
Pfx polymerase	1µL

- 22** Split reactions into 2x50µL in PCR tubes and run on a thermocycler as follows:
- 94°C for 2 min followed by 50 cycles of:
- 94°C 15 sec
- 68°C 1 min
- 23** Pool tubes containing the same DNA sample, run over QiaQuick PCR cleanup column following the manufacturer's instructions, and elute in 50µL EB.

Bind linear PCR products to beads TIMING: 1h

- 24** Resuspend streptavidin-coated beads by shaking.
- 25** Add beads (32µL/sample) to a new microcentrifuge tube (1mL maximum; use multiple tubes if greater volume is required).
- 26** Place tube on magnetic particle concentrator (MPC) for 1-2 min.
- 27** Carefully remove supernatant with a pipette.
- 28** Remove tube from MPC, add 1000µL of 1x B&W buffer; gently resuspend by pipetting.
- 29** Repeat Steps 26–28 twice for a total of 3 washes.
- 30** Remove final wash and add 2X B&W buffer (52µL/sample). Aliquot into PCR strip tubes (1 tube/sample, 50µL/tube).
- 31** Add entire volume of one sample from Step 23 to each tube.
- 32** Incubate at 25°C (room temperature) with gentle mixing for 30 min.
- 33** Place tubes on MPC for 2 min.
- 34** Carefully remove supernatant with a pipette. **CRITICAL STEP:** to avoid disturbing the beads, set the electronic multichannel pipettor to its slowest setting, place the end of the tip against the opposite side of the tube from the beads, and slowly move the tip downward as the supernatant is removed. See Supplementary Movie 1 for a demonstration.
- ?Troubleshooting
- 35** Remove tubes from MPC, add 100µL of 1x B&W buffer; gently resuspend by pipetting.

- 36** Repeat Steps 33–35 twice, but resuspend beads in 100µL LoTE buffer each time. PAUSE

POINT: samples can be stored at 4°C overnight at this point.

Second strand synthesis TIMING: 1h

- 37** Denature samples by heating in thermocycler: 95°C for 2 min, chill quickly to 4°C.
- 38** Prepare second strand mix on ice:

Component	Volume/sample
dH ₂ O	16µL
10x hexanucleotide mix	2µL
10mM dNTPS	1µL
Exo- Klenow	1µL

- 39** Collect beads with MPC, carefully discard supernatant, remove tubes from MPC, and then gently resuspend each sample in 20µL second strand mix.
- 40** Incubate samples in thermocycler at 37°C for 30 min. Mix by gently tapping tubes every 10–15 min.
- 41** Add 100 µL LoTE to each sample, collect beads in MPC, and carefully discard supernatant.
- 42** Repeat Step 41.
- 43** Resuspend beads in 100µL LoTE. PAUSE POINT: samples can be stored at 4°C overnight at this point.

MmeI digestion TIMING: 2.5h

- 44** Prepare 50µM double-stranded M12 oligonucleotide¹⁴ by combining in a new PCR tube:

M12_top (100µM in EB)	15µL
M12_bot (100µM in EB)	15µL
1M NaCl	1.5µL

- 45** Anneal oligonucleotides in a thermocycler:
- 95°C for 5 min
- Cool to 4°C at a rate of 0.1°C/sec; store in 5µL aliquots at –20°C for future use.
- 46** Prepare MmeI buffer mix on ice:

Component	Volume/sample
dH ₂ O	16.8µL
10x NEBuffer 4	2µL
32mM SAM	0.08µL

Component	Volume/sample
M12 dsDNA	0.2μL

- 47 Collect beads from Step 43 with MPC, carefully discard supernatant, remove tubes from MPC, and gently resuspend each sample in 19μL MmEI buffer mix.
- 48 Add 1μL MmEI to each sample.
- 49 Incubate in the thermocycler at 37°C for 1h. Gently mix every 10–15 min.
- 50 Add 100μL LoTE to each sample, collect beads in MPC, and then carefully discard supernatant.
- 51 Repeat Step 50.
- 52 Resuspend beads in 100μL LoTE. PAUSE POINT: samples can be stored at 4°C overnight at this point.

Linker ligation TIMING: 2.5h

- 53 Prepare 50μM stock of barcoded, double-stranded sequencing adapters (1 barcode sequence/sample) by combining in a new PCR tube:

LIB_AdaptT_(barcode) (100μM in EB)	15μL
LIB_AdaptB_(barcode) (100μM in EB)	15μL
1M NaCl	1.5μL
- 54 Anneal oligonucleotides in a thermocycler:

95°C for 5 min

Cool to 4°C at a rate of 0.1°C/sec

PAUSE POINT: Can be stored for months in 5μL single-use aliquots at –20°C for future use.
- 55 If frozen, thaw dsDNA sequencing adapters on ice (1 barcode/sample). Dilute each 50μM stock dsDNA sequencing adapter to 5μM in ice-cold 1x T4 DNA ligase buffer. CRITICAL STEP: If the adapters are not thawed on ice, they can dissociate into single-stranded DNAs
- 56 Prepare ligation mix on ice:

Component	Volume/sample
dH ₂ O	16.4μL
10x T4 DNA ligase buffer	2μL

- 57 Collect beads from Step 52 with MPC, carefully discard supernatant, remove tubes from MPC, and gently resuspend each sample in 18.4μL of ligation mix.
- 58 Add 0.6μL of 5μM dsDNA sequencing adapter (from Step 55) containing a unique barcode to each sample. Record which barcode is associated with each sample.
- 59 Add 1μL T4 DNA ligase to each sample.
- 60 Incubate in thermocycler at 16°C for 1h. Gently mix every 10–15 min.

- 61** Add 100 μ L LoTE to each sample, collect beads in MPC, and carefully discard supernatant.
- 62** Repeat Step 61.
- 63** Resuspend beads in 100 μ L LoTE. PAUSE POINT: samples can be stored at 4°C overnight at this point.

PCR and final purification TIMING: 4h

- 64** Assemble PCR mix on ice:

Component	Volume/sample
dH ₂ O	31.5 μ L
10x Pfx Buffer	10 μ L *
10mM dNTPs	2 μ L
50mM MgCl ₂	2 μ L
5 μ M LIB-PCR5	2 μ L
5 μ M LIB-PCR3	2 μ L
Pfx polymerase	0.5 μ L

* note that Pfx buffer is used at 2x standard concentration

- 65** Collect beads from Step 63 with MPC, carefully discard supernatant, remove tubes from MPC, and gently resuspend each sample in 50 μ L PCR mix on ice.
- 66** Run on a thermocycler, as follows, and prepare a 2% agarose gel (1–2 lanes/sample plus 2 ladder lanes/gel; GelGreen dye at 1:10,000 dilution; wide-tooth comb) while PCR is running:
- 94°C for 2 min followed by 18 cycles of:
- 94°C 15 sec
- 60°C 1 min
- 68°C 2 min
- Followed by:
- 68°C 4 min

PAUSE POINT: samples can be stored at 4°C overnight at this point.

- 67** Collect the beads on MPC and transfer the supernatant to new PCR tubes.
- 68** Prepare DNA ladder:
10bp DNA ladder	4 μ L
Bromophenol Blue loading dye	to 1x final concentration
dH ₂ O	to 20 μ L final volume
- 69** Add xylene cyanol loading dye to each sample (supernatant from Step 67) to 1x final concentration. CRITICAL STEP: xylene cyanol is used because bromophenol blue can migrate near the same position as the desired 125bp PCR product, obscuring the band in the gel. Bromophenol blue in the ladder lanes will provide an approximate position of the samples in the gel during electrophoresis.

- 70 Load DNA ladder onto first and last lanes of gel.
- 71 Load each sample into its own lane of the gel. If wells are too small for the full volume, split each sample over 2 lanes.
- 72 Run the gel for 30 min at 200V. CRITICAL STEP: time and voltage may need to be adjusted based on gel size and buffer composition; these parameters work well for 7cm gels run in borate buffer.
- 73 Using a non-UV based gel illuminator, excise bands at ~125bp, minimizing surplus agarose in gel fragment. Place each gel fragment in a separate microcentrifuge tube (if 2 lanes were used per sample, the similar fragments can be placed in the same tube as long as the total weight remains under 300mg). ? Troubleshooting
- 74 Clean up samples using QIAquick gel purification columns following the manufacturer's instructions (including the isopropanol step for small fragments). Be certain that all of the ethanol-containing wash buffer is removed from the column before elution by pipetting around the inner rim with a 10µL tip, if necessary. Elute in 32µL Buffer EB. Gel purification insures that primer-dimers or other erroneous products do not contribute to the DNA quantification or sequencing in subsequent steps.
- 75 Quantify each sample on QuBit or similar spectrophotometer. CRITICAL STEP: DNA concentration is typically 0.1–5ng/µL, which is below the reliable detection limit of many UV absorbance-based spectrophotometers.
- 76 All samples should subsequently be normalized to the same concentration. If all samples are ≤0.83 ng/µL, adjust an aliquot of each sample to 10nM in EB buffer. Based on an expected fragment length of 125bp, the formula for a 10nM normalization is volume of DNA (µL)=10*final volume (µL)/(DNA concentration (ng/µL)*(10⁶)*(1/649)*(1/125)). If sample concentrations are < 0.83 ng/µL, adjust an aliquot of each sample to 1nM in EB buffer. Tip: the formula for a 1nM normalization is volume of DNA (µL)=1*final volume (µL)/(DNA concentration (ng/µL)*(10⁶)*(1/649)*(1/125)).
- 77 Combine an equal volume of each sample into a single tube. CRITICAL STEP: required volumes may vary by sequencing facility. Be sure to inform the sequencing facility of whether the sample is at 10nM or 1nM.
- 78 Submit for Illumina sequencing: optimum loading concentration may vary by sequencing center or by Genome Analyzer model; we currently use 6pM for GA-IIx and HiSeq machines. ?Troubleshooting

Data analysis TIMING: Initial analysis requires approximately 3 minutes per million raw reads on a 3.06 GHz Intel Core 2 Duo processor with 4GB memory

- 79 Download the data analysis package (INSeq_analysis.zip; Supplementary Dataset 1) and sample data package (INSeq_demo.zip; downloadable from <http://gordonlab.wustl.edu/SuppData.html>) and save each to your own directory.
- 80 Bowtie is used to map sequences to the reference genome. Download this mapping software from <http://bowtie-bio.sourceforge.net/index.shtml> and install in your system. Edit the "config.txt" file in the data analysis package (INSeq_analysis) by changing the bowtie_dir variable to the path where bowtie was installed. Detailed instructions are provided in the README file in the INSeq_analysis folder (see Supplementary Dataset 1).

- 81** Go to the uncompressed data analysis package (INSeq_analysis; see Supplementary Dataset 1) and then to the "indexes" directory. From this location, create a new directory with the organism's name (e.g., BthetaVPI_5482). Put the genome file in fasta format and the annotation files (in ptt format) in the new directory. Go into the new directory and construct a bowtie index by typing the command:

```
"bowtie_dir"/bowtie-build <fasta file> <name of index, same as the new
directory>
```

?Troubleshooting

- 82** Create an experiment directory where you would like to store your analysis results. For example:

```
mkdir /Users/mwu/Documents/Experiment_1
```

- 83** Create a mapping file that contains the barcode for each sample in tab-delimited format in the experiment directory:

```
<barcode> <Sample Name>
```

- 84** From the experiment directory, run the analysis pipeline by running the wrapper script INSeq_pipeline.pl with the usage:

```
perl "path for analysis package"/INSeq_pipeline.pl -i <the raw reads file> -
m <Barcodes mapping file> -s <indexed genome name> -d
<length_disrupt_percent (max=1)>
```

Required arguments: -i gives the input raw reads file, -m gives the mapping file, -s gives the name of the indexed genome that the reads should be mapped to

Optional arguments: -d gives the region of the gene in which insertions are expected to disrupt gene function. The default is 1, which means that when the insertion falls anywhere within the gene (100%), the gene's function will be considered to be disrupted. Setting the -d argument to 0.9, for example, would exclude insertions in the distal 10% of the gene when calculating the total number of reads/insertions for that gene.

?Troubleshooting

- 85** The INSeq_pipeline.pl will generate several mappingjobsXXX.job under the experiment directory. On a desktop/laptop, run them in series by typing "sh mappingjobsXXX.job" for each job, or (on a cluster) run in parallel. Several output files will be generated as following:

- INSEQ_experiment.scarf_assigned.txt: Reads are assigned to different samples based on the sample-specific barcode.
- INSEQ_experiment.scarf.log: Some statistics for the analysis process including (1) the total number of reads, (2) the percentage of reads being mapped and trimmed (that have the transposon), (3) how many insertions in the sample (coverage) with the number of raw reads (reads after filtering in the normalization step), and (4) the scale factor being used for normalization.

Several output files will be placed in the "results" folder:

- INSEQ_experiment.scarf_Samplename.bowtiemap: Raw mapping output file from bowtie

- INSEQ_experiment.scarf_Samplename.bowtiemap_processed.txt_chromosome: Txt file containing the processed mapping output, with format

<chromosome name> <insertion position> < reads mapped to the left side of the insertion> <reads mapped to the right side of the insertion> <the total number of reads mapped to that position>

- INSEQ_experiment.scarf_Samplename.bowtiemap_processed.txt_chromosome_filt er_cpm.txt: Txt file with positions having more than 3 total reads, and normalized to counts per million reads.
- INSEQ_experiment.scarf_Samplename.bowtiemap_processed.txt_chromosome_filt er_cpm.txt_mapped: Txt file that lists genes mapped by insertions, only when the insertions located at the proximate XX percentage (specify by -d option) are considered to interrupt the function of genes. The format is

<gene name>\t<the total number of unique insertions in the gene>
\t<the sum of normalized read counts in that gene>\t<gene annotation from the ptt file>

?Troubleshooting

- 86** Clean-up. Since there are many intermediate files generated along the pipeline, if storage space is a big concern, deleting them is optional. You can run clean_up.sh from the experiment folder by using command "sh clean_up.sh", which will remove the sorted barcode assigned reads file, the folder "bcsortedseqs" and the mappingjob files.

TIMING

Steps 1–16. Isolation of crude DNA: 4h;

Steps 17–20. Clean-up of crude DNA using QIAquick columns: 1h;

Steps 21–23. Linear PCR: 2h;

Steps 24–36. Bind LPCR product to beads: 1h;

Steps 37–43. Second strand synthesis: 1h;

Steps 44– 52. MmeI digestion: 2.5h;

Steps 53–63. Linker ligation: 2.5h;

Steps 64–78. PCR and final purification: 4h;

Step 79–86. Data analysis: Initial analysis requires approximately 3 min per million raw reads on a 3.06 GHz Intel Core 2 Duo processor with 4GB memory.

Box 1. Arrayed library preparation: ~1.6h/96-well tray, plus time for picking colonies into trays and INSeq library preparation.

TROUBLESHOOTING

Troubleshooting advice can be found in Table 1.

ANTICIPATED RESULTS

The INSeq analysis pipeline produces several output files with data at the level of individual insertion locations, genes, and (optionally) operons. Good libraries should have >95% of reads that contain the transposon sequence, consistent read counts generated from both sides of each insertion, and consistent data from biological and technical replicates. A real dataset from a successful library prep is provided in the INSeq_demo files.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants from the NIH (DK30292, DK70977, and DK064540 to J.I.G; F32AI078628 and K01DK089121 to A.L.G) and by the Crohn's and Colitis Foundation of America. We thank Laura Kyro and Jonathan Eberle for their assistance in making the movies that accompany this article.

References

1. Mazurkiewicz P, Tang CM, Boone C, Holden DW. Signature-tagged mutagenesis: barcoding mutants for genome-wide screens. *Nat Rev Genet.* 2006; 7:929–939. [PubMed: 17139324]
2. Goodman AL, et al. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe.* 2009; 6:279–289. [PubMed: 19748469]
3. Erlich Y, et al. DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res.* 2009; 19
4. Prabhu S, Pe'er I. Overlapping pools for high-throughput targeted resequencing. *Genome Res.* 2009; 19:1254–1261. [PubMed: 19447964]
5. Xin X, et al. Shifted Transversal Design smart-pooling for high coverage interactome mapping. *Genome Res.* 2009; 19:1262–1269. [PubMed: 19447967]
6. Gawronski JD, Wong SM, Giannoukos G, Ward DV, Akerley BJ. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci U S A.* 2009; 106
7. Langridge GC, et al. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res.* 2009; 19:2308–2316. [PubMed: 19826075]
8. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods.* 2009; 6:767–772. [PubMed: 19767758]
9. Gallagher LA, Shendure J, Manoil C. Genome-Scale Identification of Resistance Functions in *Pseudomonas aeruginosa* Using Tn-seq. *MBio.* 2011; 2:00315–10.
10. Opijnen T, Camilli A. Genome-wide fitness and genetic interactions determined by Tn-seq, a high-throughput massively parallel sequencing method for microorganisms. *Current Protocols in Microbiology.* 2010; Chapter 1(Unit1E):3. [PubMed: 21053251]
11. Brody JR, Kern SE. Sodium boric acid: a Tris-free, cooler conductive medium for DNA electrophoresis. *Biotechniques.* 2004; 36:214–216. [PubMed: 14989083]
12. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10:R25. [PubMed: 19261174]
13. Westover BP, Buhler JD, Sonnenburg JL, Gordon JI. Operon prediction without a training set. *Bioinformatics.* 2005; 21:880–888. [PubMed: 15539453]
14. Morgan RD, Bhatia TK, Lovasco L, Davis TB. MmeI: a minimal Type II restriction-modification system that only modifies one DNA strand for host protection. *Nucleic Acids Res.* 2008; 36:6558–6570. [PubMed: 18931376]

(a) Extract DNA from samples (steps 1-20)

(b) Linear PCR with biotinylated primer (steps 21-23)

(c) Bind to beads, fill-in (steps 24-43)

(d) Digestion, ligation (steps 44-63)

(e) Amplification, sequencing (steps 64-78)

Figure 1. Overview of insertion sequencing (INSeq) protocol

(i) Genomic DNA is extracted from input transposon mutant libraries or from output samples, e.g., fecal samples obtained from gnotobiotic mice after colonization. The procedure described in Steps 1–16 can be used for extracting DNA from the input mutant population directly or the output populations. (ii) Linear PCR is performed to amplify the transposon with adjacent chromosomal DNA. (iii) The PCR product is purified by binding to solid-phase beads; second strand DNA is synthesized. (iv) DNA fragments are digested by MmeI and then ligated to sequencing adapters with a sample-specific barcode. (v) Products from stage (iv) are amplified in a limited PCR step and sequenced. Inverted repeat (IR) region is shown as a black triangle. For convenience, stages (iii)–(v) illustrate only one side of the transposon; in reality both sides are amplified and sequenced.

Table 1

Troubleshooting.

Step	Problem	Possible reason	Solution
34	Beads are lost during wash steps	Electronic multichannel pipettor is too fast	If alternate pipettors are not available, perform this step manually
73	Bands are faint or inconsistent between samples	Ligation adapters not properly annealed, or have barcodes that prevent proper annealing	Use technical replicates (same input DNA with different sample-specific barcodes) to determine whether the barcode or the input DNA is the problem.
78	Illumina sequencing produces fewer reads than expected	Sequence characteristics prevent proper function of Illumina base-calling software	Provide sequencing facility with information about common and variable bases in your expected sequences based on the barcodes used
81	Error message "bowtie-build: No such file or directory"	Command did not specify the location of bowtie	Make sure to include the entire path for bowtie, e.g. <code>"/Users/XXX/bowtie-0.12.7/bowtie-build Bt.fna Bt.VPI 5482"</code> .
84	Analysis pipeline cannot recognize the raw reads file	The raw reads file is not in the scarf format.	Convert your raw reads file to scarf format.
85	Error message "bowtie: No such file or directory"	The analysis package could not locate bowtie	Check the "config.txt" in Step 80; make sure the bowtie_dir variable is the path where your bowtie installation is located. e.g., <code>bowtie_dir="/Users/XXX/bowtie-0.12.7"</code> .
85	Error message "Could not locate a bowtie index corresponding to basename 'XXX'"	The analysis package could not find the bowtie index. The analysis package is looking for the index file located in "INSEQ_analysis/indexes/XXX/XXX" (XXX is the "indexed genome name" specified in step 84 by argument <code>-s</code>)	Check the location and the name of your index file. Make sure it is UNDER the "INSEQ_analysis/indexes" directory (should be in the position exactly inside of the "indexes" folder), and check that the name of the index, which is the prefix of the .X.ebwt file, is consistent with the name of the directory. If not, re-make the ebwt file using bowtie-build and repeat step 81.
85	The gene mapped results show all genes have zero insertions while the insertion results show that there are several insertions along the genome.	The analysis package matches the gene annotation file (.ptt) to chromosome (.fna) by names. If the naming is not consistent, the pipeline cannot locate the insertions to genes correctly.	Remove any spaces in any names, and re-name the header names of the .fna files to match the .ptt files, or rename the .ptt files to match the header names of the .fna files.