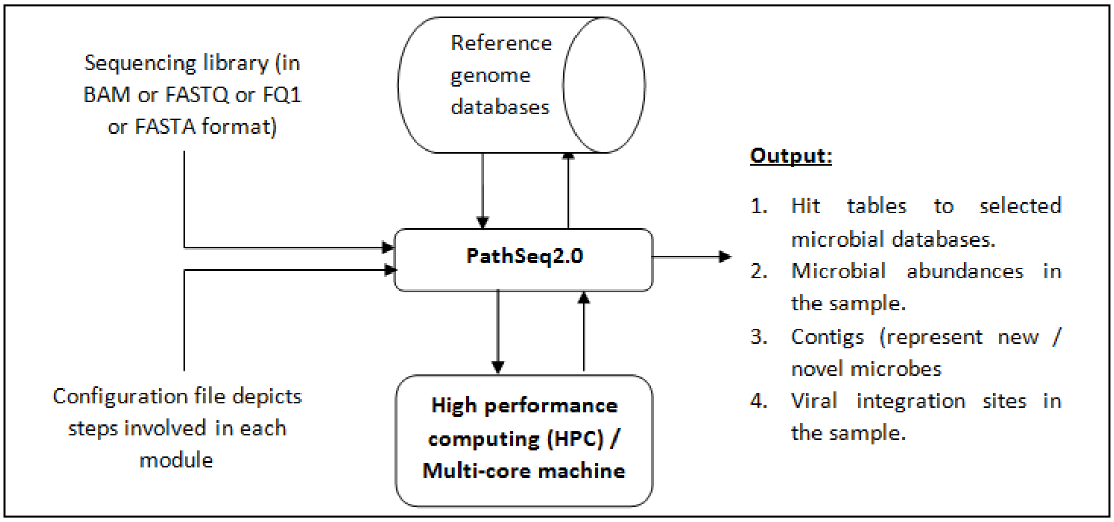
README

We present a customizable computational tool, PathSeq2.0, to enable the discovery and identification of microbial sequences in metagenomic samples. This is a flexible, user-adaptable, pipeline that is easily assembled using a configuration file tailored towards particular sequencing experiment parameters such as type of sequencing library, sample of origin, sequencing technology, host species and library read length. PathSeq2.0 also provides auxiliary tools that facilitate the identification of viral integration sites within the host genome and quantification of microbial abundance in the sample set. This package is available at https://github.com/ChandraPedamallu/PathSeq.

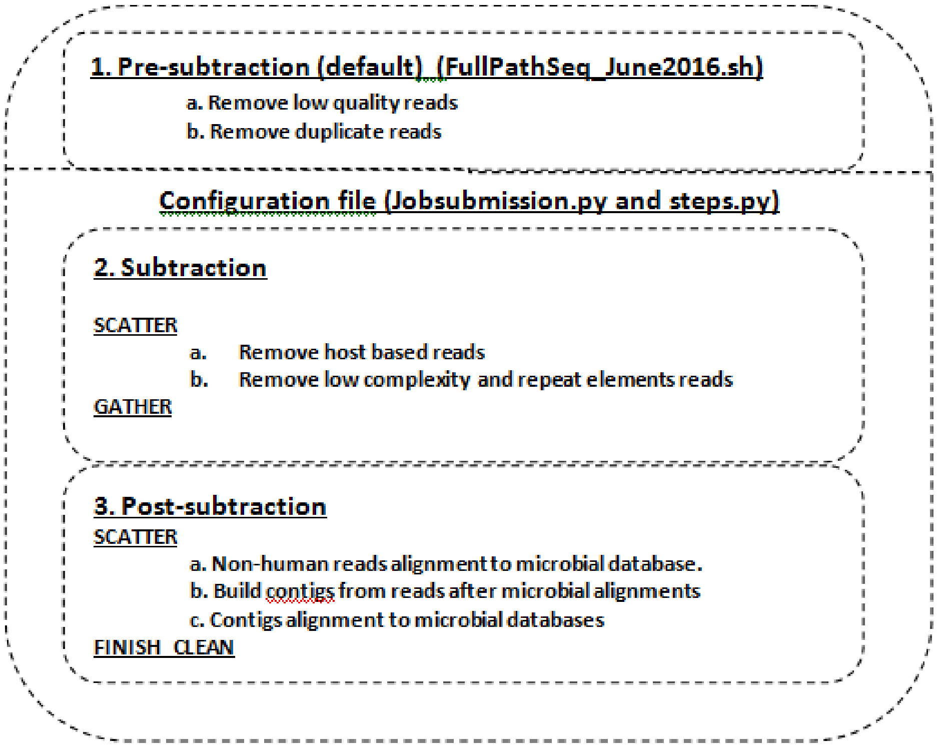
1) Schematic diagram of PathSeq2.0 pipeline.

(a) High level design diagram of PathSeq2.0.



Sequencing library (in BAM or FASTQ or FQ1 or FASTA format)
Reference genome databases
PathSeq2.0
Configuration file depicts steps involved in each module
High performance computing (HPC) / Multi-core machine
Output:
1. Hit tables to selected microbial databases.
2. Microbial abundances in the sample.
3. Contigs (represent new / novel microbes
4. Viral integration sites in the sample.

(b) Steps involved in PathSeq2.0 and associated generic configuration file.



1. Pre-subtraction (default) (FullPathSeq_June2016.sh)

a. Remove low quality reads
b. Remove duplicate reads

Configuration file (Jobsubmission.py and steps.py)

2. Subtraction

SCATTER

a. Remove host based reads
b. Remove low complexity and repeat elements reads

GATHER

3. Post-subtraction

SCATTER

a. Non-human reads alignment to microbial database.
b. Build contigs from reads after microbial alignments
c. Contigs alignment to microbial databases

FINISH CLEAN

2) Dependencies package

(a) PathSeq2.0 dependencies package:

Table below indicates PathSeq2.0 dependency software's and its associated versions. It also indicates whether the dependency software is "M"- Mandatory and "P"-Preferred and location for download.

Software	Version	M/P	Location
Java	1.7	M	http://www.oracle.com/technetwork/indexes/downloads/index.html#java
Python	2.7 or 2.7.1	M	https://www.python.org/
BLAST	2.2.30+	M	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.30/
Repeat Masker	4.0.5	P	http://www.repeatmasker.org/RMDownload.html
BWA	0.7.12	M	https://sourceforge.net/projects/bio-bwa/files/
Velvet	1.2.10	M	http://www.ebi.ac.uk/~zerbino/velvet/
Picard (Mergesamfiles.jar)	1.26	M	Already included under the "3rdparty" folder

SAM tools (sam-1.35.jar and sam-1.52.jar)	1.35 and 1.52	M	Already included under the “3rdparty” folder

(b) PathSeq2.0 integration event finder:

Table indicates PathSeq2.0 integration event finder dependency software’s and its associated versions. It also indicates whether the dependency software is “M”- Mandatory and “P”- Preferred and location for download.

Software	Version	M / P	Location
Java	1.7	M	http://www.oracle.com/technetwork/indexes/downloads/index.html#java
Samtools	0.1.19	M	https://sourceforge.net/projects/samtools/files/samtools/
Tophat	2.0.14	M	https://ccb.jhu.edu/software/tophat/index.shtml
SamBlaster	0.1.22	M	https://github.com/GregoryFaust/samblaster (Already included under the “Integration_Events /software” folder)
BWA	0.7.12	M	https://sourceforge.net/projects/bio-bwa/files/
LUMPY-SV	1.2.10	M	https://github.com/arq5x/lumpy-sv (Already included under the “Integration_Events /software” folder)

3) PathSeq2.0 databases:

The following are the reference databases used in a typical PathSeq2.0 run.

Sl. No.	Database name	Detailed description
1	**Ensembl database**	This database is combination of cDNA and RNA data from Ensembl and NCBI. BWA and BLAST compatible databases are built and primarily used in substraction module. Moreover, it is used for RNASeq sequencing library analysis.
2	**Female Genome**	This database is downloaded from 1000 genomes. BWA and BLAST compatible databases are built and primarily used in substraction module. Moreover, it is used for DNASeq sequencing library analysis.

Sl. No.	Database name	Detailed description
3	**Genome plus transcriptome database**	This database is downloaded from NCBI blast db directory. BWA and BLAST compatible databases are built and primarily used in substraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.
4	**Human Reference genome**	This database is downloaded from NCBI blast db directory. BWA and BLAST compatible databases are built and primarily used in substraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.
5	**Bacterial genomes**	This database is downloaded from NCBI reference genomes. Plasmid, Private and contig sequences are removed from the downloaded files. BLAST compatible database is built and primarily used in post-subtraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.
6	**Viral genomes**	This database is downloaded from NCBI reference genomes. Private and patent sequences are removed from the downloaded files. BLAST compatible database is built and primarily used in post-subtraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.
7	**Archeae genomes**	This database is downloaded from NCBI reference genomes. Plasmid, Private and contig sequences are removed from the downloaded files. BLAST compatible database is built and primarily used in post-subtraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.
8	**Fungi genomes**	This database is downloaded from NCBI reference genomes. Plasmid, Private and contig sequences are removed from the downloaded files. BLAST compatible database is built and primarily used in post-subtraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.
9	**Phage genomes**	This database is downloaded from NCBI reference genomes. Private and patent sequences are removed from the downloaded files. BLAST compatible database is built and primarily used in post-subtraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.

Sl. No.	Database name	Detailed description
10	**Human plus microbial genomes**	This is combination of Serial numbers 4 to 9. BLAST compatible database is built and primarily used in post-subtraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.
11	**Premegablast database**	This is a combination ribosomal, and mitochondrial sequences from the host. BLAST compatible database is built and primarily used in post-subtraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.
12	**16S and 23S database**	This is a combination 16S and 23S bacterial sequence. BLAST compatible database is built and primarily used in post-subtraction module. Moreover, it is used for DNASeq / RNASeq sequencing library analysis.

4) Configuration file

(a) Detailed description

Configuration file outlines the steps involved in PathSeq2.0 modules, which are organized into “SCATTER-GATHER” blocks. Each starts with “SCATTER” and ends with one of the “GATHER / GATHERASSEMBLER / FINISH_CLEAN” keywords. In the case of HPC mode, each “SCATTER” keyword is followed by parameters associated with the job such as such as chunk size (number of reads to be analyzed in a single job), number of threads used to run tools in individual steps, and cluster specific job submission prefix command (Example. In the case of UGER and LSF cluster specific prefix commands are “qsub -q queue_name -l m_mem_free=8g” and “bsub -q queue_name -R "rusage[mem=1]"” respectively. However, in the case of STANDALONE mode chunk size and cluster specific job submission prefix command are left empty except the number of threads. The “GATHER” keyword gathers the scattered jobs before the start of the next “SCATTER-GATHER” block. “GATHERASSEMBLER” is specific for the gather step after velvet assembler run. The “FINISH_CLEAN” is specific for cleaning the intermediate files and create final output files under “Final_combine_results” directory.

Each “SCATTER – GATHER” block contains steps to be executed and which starts with the tool to run (Valid tools are BWA or BLAST (MEGABLAST / BLASTN / BLASTX / TBLASTX) or REPETMASKER or VELVET) and the location of reference database (Eg. Bacterial Genomes or Viral genomes or Human reference genome or etc.). The default alignment score to call a read mapped or not by BLAST aligners is <=1E-7 and BWA uses the default mapping score.

(b) Recommended configuration files

a. total RNA / mRNA sequencing library

```
SCATTER:1000000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
BWA:Location of Ensembl human database:Reads mapped to Ensembl human database
BWA:Location of Human reference genome:Reads mapped to Human reference database
GATHER
SCATTER:500000:1:qsub -q long -l m_mem_free=8g(Univa Grid Engine specific)
REPEATMASKER
MEGABLAST:Location of Ensembl human database:Reads mapped to Ensembl human database
MEGABLAST:Location of Human reference genome:Reads mapped to Human Reference database
GATHER
SCATTER:100000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
PREMEGABLAST:Location of Human mitochondrial and ribosomal database:Reads to Human mitochondrial and ribosomal database
MEGABLAST:Location of Phage database:Reads mapped to Phage database
MEGABLAST:Location of Human reference genome + Microbial genome database:Reads mapped to HUMAN + MICROBES database
GATHER
SCATTER:NA:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
VELVET:SINGLEEND:::Building Contigs using single end assembly
GATHERASSEMBLER
SCATTER:10000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
MEGABLAST:Location of Human reference genome + Microbial genome database:Contigs mapped to HUMAN + MICROBES database
FINISH_CLEAN
```

b. Whole genome sequencing (WGS) / Whole exome sequencing library (WES)

```
SCATTER:1000000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
BWA:Location of Female reference genome:Reads mapped to Female reference genome
BWA:Location of Human reference genome:Reads mapped to Human reference genome GATHER
SCATTER:500000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
REPEATMASKER
MEGABLAST:Location of Ensembl human database:Reads mapped to Ensembl human database
MEGABLAST:Location of Human reference genome:Reads mapped to Human Reference Genome
GATHER
SCATTER:100000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
PREMEGABLAST:Location of Human mitochondrial and ribosomal database:Reads to Human mitochondrial and ribosomal database
MEGABLAST:Location of Phage database:Reads mapped to Phage database
MEGABLAST:Location of Human reference genome + Microbial genome database:Reads mapped to HUMAN + MICROBES database
GATHER
SCATTER:NA:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
VELVET:SINGLEEND:::Building Contigs using single end assembly
GATHERASSEMBLER
SCATTER:10000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
MEGABLAST:Location of Human reference genome + Microbial genome database:Contigs mapped to HUMAN + MICROBES database
```

```
FINISH_CLEAN
```

c. Library generated from stool samples

```
SCATTER:100000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
PREMEGABLAST:Location of Human mitochondrial and ribosomal database:Reads to Human mitochondrial and ribosomal database
MEGABLAST:Location of Phage database:Reads mapped to Phage database
MEGABLAST:Location of Human reference genome + Microbial genome database:Reads mapped to HUMAN + MICROBES database
GATHER
SCATTER:NA:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
VELVET:SINGLEEND:::Building Contigs using single end assembly
GATHERASSEMBLER
SCATTER:10000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
MEGABLAST:Location of Human reference genome + Microbial genome database:Contigs mapped to HUMAN + MICROBES database
FINISH_CLEAN
```

d. Library generated from 16S experiment

```
SCATTER:10000:1:qsub -q long -l m_mem_free=8g (Univa Grid Engine specific)
MEGABLAST:Location of bacterial 16s database:Reads mapped bacterial 16S database
FINISH_CLEAN
```

5) Typical command line and Results

(a) PathSeq2.0:

Typical command line for PathSeq2.0 command as follows.

```
< PathSeq2.0 installation directory>/FullPathSeq_June2016.sh <Location of Input file> <Type of Input file: BAM/FASTQ/FASTA/FQ1> <Location of configuration file > <Mode:UGER/STANDALONE/LSF>
```

Final results from PathSeq2.0 are listed under “Final_combine_results” directory. Unmapped reads after each “SCATTER-GATHER” block is listed with prefix of “qf_1.unique.fq1” and file ends with “.unmappedfinal.fq1”. In addition, for each block that is processed on reads there is “unmappedfinal.fq1” appended. However, contigs generated from still-unmapped reads (reads remaining after mapping to host and known microbial databases) file ends with “.contigs.fq1” and unmapped contig file also ends with “.unmappedfinal.fq1” and follows similar annotation as above.

Alignment hits are assembled into “Hittables” which are tab-delimited files contains read and contig level information and its assignments to the reference databases. Each “Hittable” prefixed with the tool that is used for run and ends with “SCATTER-BLOCK count_(Step count-1)”. For user friendliness, “REPORT.HTML” provides number of reads after each step and also provides links to each file.

(b) PathSeq2.0 integration event finder

PathSeq2.0 integration event finder is a tool that identifies microbial (typically virus) insertion site in the host genome using RNASeq and DNASeq sequencing libraries. This tool make use of the “Hittable” (generated through PathSeq2.0 mapping of reads to relevant microbial database), still-unmapped reads, sequencing library in BAM format and combined genome (host and microbe of interest genomes). The sequencing library in BAM format is taken for extracting read pairs based on the names of the sequences mapped to microbe of interest and still-unmapped reads. PathSeq2.0 integration event finder employs Tophat2.0.14 (enabled fusion detection) and (BWA plus SAMBLASTER plus LUMPY-SV) algorithms for RNASeq and DNASeq libraries respectively.

The recommended number of spanning reads and flanking reads supporting an integration event is at least 3 and 10 respectively. Flanking read pairs were defined as having one end of the paired end read mapped to the microbe genome and its mate pair mapped to the host genome. Spanning reads were defined as having one end of the paired end read spanning the integration junction and its mate pair mapped to either the host or microbe genome. Integration events from RNASeq and DNASeq are in tab-delimited file and variant call file format respectively. Each output file contains not only the integration events between the host and microbe events but also some of the events within host-host or microbe-microbe events.

Typical command line for PathSeq2.0 integration event command as follows:
< PathSeq2.0 installation directory>/ Integration_Events/Integration_Site.sh <Hittable that contains viral reads of interest> < Still-unmapped reads in FQ1> <Database that contains human and viral genome of interest> <Location of original BAM file> <RNASEQ/DNASEQ> <Read length> DEFAULT

(c) Bacterial abundance quantification tool

In addition to integration event finder, PathSeq2.0 also contains a bacterial abundance quantification tool. This tool makes use of “Hittable” generated by reads mapping to relevant microbial database, number of host reads in the sample and full bacterial taxonomy tree. Bacterial abundance quantification tool calculates abundances for each sample at different taxa levels (includes but not limited to Phylum, Family, Order, Genus, Species). This tool considers only reads that are mapped at >=90% identity and >=90% query coverage. In case of read mapped to multiple subjects at same mapping quality (i.e. >=90% identity and >=90% query coverage), we equally distribute the read (i.e. if a read maps equally good to x subjects, then assigned value to each subjects is 1/x). The abundance metric of microbe 1 in sample 1 is calculated as (Number of reads mapped to a microbe 1/Number of host reads in sample 1)*10^6.

Typical command line for PathSeq2.0 bacterial abundance command as follows:

```
java –classpath < PathSeq2.0 installation directory>/ Abundance_Bacteria/
RAM_Rawreadscounts  <Output file> <names.dmp location> <nodes.dmp location> <Tab-
delimited file that contains sample name, location of hittables associated to the sample and
number of host reads in the associated sample>
```