SeqHBase: a big data toolset for family-based sequencing data analysis

Supplementary material

1. Introduction

SeqHBase is a big data toolset developed based on Apache Hadoop (Apache Foundation, Hadoop) and HBase (Apache Foundation, HBase) infrastructure. It is designed for analysing family-based sequencing data to detect *de novo*, inherited homozygous or compound heterozygous mutations. The toolset is efficient, reliable, and capable of handling large and/or extended family-based sequencing data for analysis.

SeqHBase provides a number of features and functionalities. It extracts variant, variation, and coverage (read-depth) information from three commonly used file formats, including tab (or comma) delimited variant annotated files (e.g. CSV files), VCF files, and compressed BAM files.

This supplementary material consists of four parts. The first part describes installation and dependencies. The second part discusses loading sequencing data into HBase. The third part describes detecting *de novo*, inherited homozygous or compound heterozygous mutations using SeqHBase. In the final section, the mutations detected, when analysing three different family-based sequencing data sets, are shown in more detail.

2. Installation

As SeqHBase builds on top of Apache Hadoop and HBase, which itself relies on Hadoop and HBase for job execution, the installation requires a working Hadoop and HBase setup, such as Amazon's Elastic MapReduce service or an in-house Hadoop and HBase cluster. For more information about setting up an Apache Hadoop and HBase cluster, please refer to http://hadoop.apache.org/ and http://hbase.apache.org/.

2.1 Dependencies

We developed and tested SeqHBase with the following dependencies:

- Hadoop version 1.2.1
- HBase version 0.94.19

2.2 Environment variables

- Set Hadoop-related variables (e.g., HADOOP_HOME) for your installation
- Set HBASE_HOME to point to your HBase installation
- Set JAVA_HOME to point to your Java installation

2.3 Installing a pre-compiled release

1. Download the latest (current 1.00) SeqHBase release from http://seqhbase.omicspace.org/.

2. Untar the release into an installation directory of your choice; e.g.,

```
$ tar zxvf seqhbase_1.00.tar.gz
```

3. ETL sequencing data into your Hadoop and HBase cluster

SeqHBase efficiently extracts, transforms, and loads (ETL) your variant, variation, and coverage information from three types of files as described in Table 1 of the manuscript.

3.1 **Variant**: annotated variant files generated by ANNOVAR (Wang et al., 2010) are used to extract variant information, including chromosome number, start position, end position, reference allele, alternative allele, frequency in the 1000 Genome Project (Abecasis et al., 2012) and/or the 6500 EPS project (Exome Variant Server, 2014), ClinVar (Landrum et al., 2014), CADD score (Kircher et al., 2014), biological function, and multiple diverse function-relevant scores, such as PolyPhen-2 score (Adzhubei et al., 2010), SIFT score (Kumar et al., 2009), and others. When developing SeqHBase, we applied ANNOVAR to annotate variants in sequencing data sets. However, annotated information generated by other annotation programs, such as SnpEff (http://snpeff.sourceforge.net/), can also be applied in SeqHBase.

```
$ seqhbase.sh --memory 1024 --csv-file $ANNOTATED_FILE.csv --sample-id
$FAM_ID:$IND_ID
```

where `$FAM_ID` is family ID while `$IND_ID` is individual ID.

3.2 **Variation**: VCF files are used to extract variation information, including sample family ID, individual ID, called variant genotypes, coverage (read-depth), and Phred quality scores.

```
$ seqhbase.sh --memory 1024 --vcf-file $VCF_FILE.vcf --sample-id
$FAM_ID:$IND_ID
```

3.3 **Coverage (read-depth)**: BAM files are used to extract data regarding coverage of each site of every sequencing sample (~3 billion sites in a WGS data). In downstream analyses, the read-depth information can identify if no-call sites are reference-consistent with high quality or reference-inconsistent caused by low quality. A specific function is developed for quickly generating the read depths of each site from BAM files, similar to SAMtools (Li et al., 2009) pileup function. SeqHBase also supports loading coverage information into HBase from a pileup file generated by SAMtools.

```
$ seqhbase.sh --memory 4096 --pileup-file $BAM_FILE.bam --sample-id
$FAM_ID:$IND_ID
```

Or

```
$ seqhbase.sh --memory 1024 --pileup-file $PILEUP_FILE.gz --sample-id
$FAM_ID:$IND_ID
```

4. Detection of *de novo*, inherited homozygous or compound heterozygous mutations

4.1 Commands

***De novo* and autosomal recessive (or X-linked) screens** - command line as follows:

```
$ seqhbase.sh --memory 1024 --ped-file $PEDFILE.ped --list-denovo --
min-coverage-screen 20 --maf 0.01 --func-list $FUNCLIST –-exonic-func-
list $EXONICFUNCLIST --query --out $OUTPUT
```

Compound het screens - command line as follows:

```
$ seqhbase.sh --memory 1024 --ped-file $PEDFILE.ped --list-comp-het --
min-coverage-screen 20 --maf 0.01 --func-list $FUNCLIST –-exonic-func-
list $EXONICFUNCLIST --query --out $OUTPU
```

By default, `$FUNCLIST` is "`exonic,splicing`" and `$EXONICFUNCLIST` is started with any of the "`nonsynonymous,stopgain,stoploss,frameshift`".

According to ANNOVAR, the `$FUNCLIST` can be one or more of the following values:

Value	Default precedence	Explanation
exonic	1	variant overlaps a coding exon
splicing	1	variant is within 2-bp of a splicing junction (use -splicing_threshold to change this)
ncRNA	2	variant overlaps a transcript without coding annotation in the gene definition
UTR5	3	variant overlaps a 5' untranslated region
UTR3	3	variant overlaps a 3' untranslated region
intronic	4	variant overlaps an intron
upstream	5	variant overlaps 1-kb region upstream of transcription start site
downstream	5	variant overlaps 1-kb region downstream of transcription end site (use -neargene to change this)
intergenic	6	variant is in an intergenic region

According to ANNOVAR, the `$EXONICFUNCLIST` can be one or more of the following annotations:

Annotation	Precedence	Explanation
frameshift insertion	1	an insertion of one or more nucleotides that causes frameshift changes in a protein coding sequence
frameshift deletion	2	a deletion of one or more nucleotides that causes frameshift changes in a protein coding sequence

frameshift block substitution	3	a block substitution of one or more nucleotides that causes frameshift changes in a protein coding sequence
stopgain	4	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that leads to the immediate creation of a stop codon at the variant site. For frameshift mutations, the creation of a stop codon downstream of the variant will not be counted as "stopgain"
stoploss	5	a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that leads to the immediate elimination of a stop codon at the variant site
nonframeshift insertion	6	an insertion of 3 or multiples of 3 nucleotides that does not cause frameshift changes in a protein coding sequence
nonframeshift deletion	7	a deletion of 3 or multiples of 3 nucleotides that does not cause frameshift changes in a protein coding sequence
nonframeshift block substitution	8	a block substitution of one or more nucleotides that does not cause frameshift changes in a protein coding sequence
nonsynonymous SNV	9	a single nucleotide change that causes an amino acid change
synonymous SNV	10	a single nucleotide change that does not cause an amino acid change
Unknown	11	unknown function (due to various errors in the gene structure definition in the database file)

4.2 Some input parameters for detecting mutations

- **`--list-denovo`**: list *de novo* and autosomal recessive (or X-linked) mutations in nuclear families.
- **`--list-comp-het`**: list compound heterozygous mutations in nuclear families.
- **`--min-coverage-screen` {20} [optional]**: specify a minimum coverage (read depth) for screens; the default value is 20.
- **`--maf` {0.01} [optional]**: specify a maximum variant allele frequency in 1000 Genome Project and Exome Project Server; the default value is 0.01.
- **`--out` {foldername & fileroot} [optional]**: specify output foldername and output root filename.

For more detailed input parameters, please refer to the SeqHBase website http://seqhbase.omicspace.org/.

5. Results of analysing three different types of sequencing data sets

5.1 Analysis of a whole genome sequencing data set on a 5-member nuclear family

Table S1. Variants conforming to the disease models of *de novo*, inherited homozygous or compound heterozygous mutations carried by the affected in the 5-member nuclear family based on the analysis criteria of each variant to have a low population frequency (MAF ≤ 1%), read depth ≥ 20X and to have been annotated as being “nonsynonymous,” “stop-gain,” “stop-loss,” “splicing”, or “frame-shift” changes. In the following table, one most plausible *de novo* mutation (chr1:149898811C>A) and one possible compound heterozygous mutation in *CAND2* could be associated with the syndromes studied.

Disease model	Location	Ref	Alt	Gene	Function
De novo	1:149898811	C	T	SF3B4 (NM_005850:exon4:c.164-1G>A)	splicing
De novo	3:10114944	A	C	FANCD2(NM_001018115:exon28:c.A2613C:p.K871N, NM_033084:exon28:c.A2613C:p.K871N)	nonsynonymous
De novo	3:75714971	C	A	FRG2C (NM_001124759:exon4:c.C628A:p.L210M)	nonsynonymous
De novo	5:175528119	T	C	FAM153B(NM_001265615:exon10:c.T401C:p.M134T)	nonsynonymous
De novo	9:139902962	C	T	ABCA2 (NM_001606:exon48:c.G7181A:p.R2394Q, NM_212533:exon48:c.G7271A:p.R2424Q)	nonsynonymous
Autosomal recessive	11:5989415	C	T	OR56A5 (NM_001146033:exon1:c.G310T:p.V104L)	nonsynonymous
De novo	22:25011077	C	T	GGT1 (NM_001032364:exon7:c.C365T:p.S122L, NM_001032365:exon7:c.C365T:p.S122L,GGT1:NM)	nonsynonymous
Comp het	3:12854873	A	T	CAND2 (NM_012298:exon5:c.A713T:p.E238V, NM_001162499:exon7:c.A992T:p.E331V)	nonsynonymous
	3:12858158	C	T	CAND2 (NM_012298:exon8:c.C1448T:p.A483V, NM_001162499:exon10:c.C1727T:p.A576V)	nonsynonymous
Comp het	9:95272297	C	T	ECM2 (NM_001197295:exon6:c.G1124A:p.R375H, NM_001197296:exon6:c.G1124A:p.R375H, NM_001393:exon6:c.G1190A:p.R397H)	nonsynonymous
	9:95277059	C	T	ECM2 (NM_001197295:exon4:c.G842A:p.R281Q, NM_001197296:exon4:c.G842A:p.R281Q, NM_001393:exon4:c.G908A:p.R303Q)	nonsynonymous

5.2 Analysis of a whole exome sequencing data set on a 4-member nuclear family

Table S2. Variants conforming to the disease models of *de novo*, or compound heterozygous mutations carried by the affected in the 4-member nuclear family based on the analysis criteria

of each variant to have a low population frequency (MAF ≤ 1%), read depth ≥ 20X and to have been annotated as being “nonsynonymous,” “stop-gain,” “stop-loss,” “splicing”, or “frame-shift” changes. In the following table, the most plausible compound heterozygous mutations in *PKLR* could be associated with the syndrome studied.

Disease model	Location	Ref	Alt	Gene	Function
De novo	1:151337690	G	C	SELENBP1(NM_001258288:exon9:c.C926G:p.S309C, NM_001258289:exon10:c.C1238G:p.S413C, NM_003944:exon10:c.C1112G:p.S371C)	nonsynonymous
De novo	1:202731846	G	T	KDM5B (NM_006618:exon7:c.C899A:p.S300Y)	nonsynonymous
De novo	3:156413682	A	C	TIPARP (NM_001184717:exon4:c.A1115C:p.N372T, NM_001184718:exon4:c.A1115C:p.N372T, NM_015508:exon4:c.A1115C:p.N372T)	nonsynonymous
De novo	4:57839405	T	C	NOA1 (NM_032313:exon3:c.A1424G:p.H475R)	nonsynonymous
De novo	5:177546780	A	G	N4BP3 (NM_015111:exon2:c.A196G:p.S66G)	nonsynonymous
De novo	6:136590698	C	T	BCLAF1 (NM_001077440:exon9:c.G2090A:p.R697H, NM_001077441:exon9:c.G1577A:p.R526H, NM_014739:exon9:c.G2096A:p.R699H)	nonsynonymous
De novo	7:93540153	G	C	GNGT1 (NM_021955:exon3:c.G148C:p.E50Q)	nonsynonymous
De novo	8:17611809	G	C	MTUS1 (NM_001001924:exon2:c.C1508G:p.P503R, NM_001001925:exon2:c.C1508G:p.P503R)	nonsynonymous
De novo	12:39726179	T	C	KIF21A (NM_001173465:exon19:c.A2780G:p.K927R, NM_001173463:exon20:c.A2849G:p.K950R, NM_017641:exon20:c.A2849G:p.K950R, NM_001173464:exon21:c.A2888G:p.K963R)	nonsynonymous
De novo	14:64954410	G	A	ZBTB25 (NM_006977:exon3:c.C539T:p.T180I)	nonsynonymous
De novo	14:67671483	A	C	FAM71D (NM_173526:exon5:c.A589C:p.T197P)	nonsynonymous
De novo	15:92690379	A	G	SLCO3A1 (NM_001145044:exon8:c.A1678G:p.I560V, NM_013272:exon8:c.A1678G:p.I560V)	nonsynonymous
De novo	17:21319208	C	T	KCNJ18 (NM_001194958:exon3:c.C554T:p.A185V), KCNJ12 (NM_021012:exon3:c.C554T:p.A185V)	nonsynonymous
De novo	17:7849130	G	A	CNTROB(NM_001037144:exon13:c.G1819A:p.V607M, NM_053051:exon13:c.G1819A:p.V607M)	nonsynonymous

De novo	19:3179427	G	A	S1PR4 (NM_003775:exon1:c.G637A:p.A213T)	nonsynonymous
De novo	21:10942923	G	A	TPTE (NM_199260:exon10:c.C550T:p.R184W, NM_199259:exon11:c.C610T:p.R204W, NM_199261:exon12:c.C664T:p.R222W)	nonsynonymous
Comp het	1:155260382	G	A	PKLR (NM_000298:exon11:c.G1706A:p.R569Q, NM_181871:exon11:c.G1613A:p.R538Q)	nonsynonymous
	1:155264120	C	G	PKLR (NM_000298:exon7:c.G1022C:p.G341A, NM_181871:exon7:c.G929C:p.G310A)	nonsynonymous
Comp het	14:72054708	G	A	SIPA1L1 (NM_001284247:exon1:c.G119A:p.R40Q, NM_001284245:exon2:c.G119A:p.R40Q, NM_001284246:exon2:c.G119A:p.R40Q, NM_015556:exon2:c.G119A:p.R40Q)	nonsynonymous
	14:72055745	A	G	SIPA1L1 (NM_001284247:exon1:c.A1156G:p.M386V, NM_001284245:exon2:c.A1156G:p.M386V, NM_001284246:exon2:c.A1156G:p.M386V, NM_015556:exon2:c.A1156G:p.M386V)	nonsynonymous

5.3 Analysis of a whole genome sequencing data set on a 10-member 3-generation family

Table S3. Variants conforming to the disease models of *de novo*, inherited homozygous or compound heterozygous mutations carried by both of the two affecteds and/or their mother in the 10-member 3-generation family based on the analysis criteria of each variant to have a low population frequency (MAF ≤ 1%), read depth ≥ 20X and to have been annotated as being “nonsynonymous,” “stop-gain,” “stop-loss,” “splicing”, or “frame-shift” changes. In the following table, the X-linked mutation was a *de novo* mutation carried by the mother of the two affecteds; and the mutation was inherited by both of the two affecteds. Two compound heterozygous mutations located in one gene (*SYNE2*) were carried by the two affecteds individually.

Disease model	**Location**	**Ref**	**Alt**	**Gene**	**Function**
De novo	1:183209311	C	T	LAMC2 (NM_005562:exon21:c.C3206T:p.T1069M, NM_018891:exon21:c.C3206T:p.T1069M)	nonsynonymous
De novo	1:38168955	G	A	CDCA8 (NM_001256875:exon7:c.G520A:p.V174M, NM_018101:exon8:c.G520A:p.V174M)	nonsynonymous
De novo	2:70408444	G	A	C2orf42 (NM_017880:exon3:c.C674T:p.S225F)	nonsynonymous
De novo	2:97370322	C	A	FER1L5(NM_001113382:exon52:c.C6175A:p.Q2059K)	nonsynonymous

De novo	3:169557891	G	A	LRRC31 (NM_001277128:exon8:c.C1370T:p.P457L, NM_024727:exon9:c.C1538T:p.P513L)	nonsynonymous
De novo	4:6279312	C	A	WFS1 (NM_001145853:exon2:c.C130A:p.P44T, NM_006005:exon2:c.C130A:p.P44T)	nonsynonymous
De novo	6:94068112	A	G	EPHA7 (NM_001288629:exon4:c.T850C:p.Y284H, NM_004440:exon4:c.T850C:p.Y284H)	nonsynonymous
De novo	8:2017399	C	T	MYOM2 (NM_003970:exon7:c.C656T:p.A219V)	nonsynonymous
De novo	11:124253170	G	A	OR8B2 (NM_001005468:exon1:c.C70T:p.R24W)	nonsynonymous
De novo	11:72423290	C	T	ARAP1 (NM_001135190:exon5:c.G238A:p.V80I, NM_015242:exon5:c.G238A:p.V80I, NM_001040118:exon7:c.G973A:p.V325I)	nonsynonymous
De novo	12:44148180	C	T	PUS7L (NM_001098614:exon2:c.G869A:p.R290K, NM_001098615:exon2:c.G869A:p.R290K, NM_031292:exon2:c.G869A:p.R290K)	nonsynonymous
De novo	13:33680994	A	G	STARD13 (NM_052851:exon13:c.T2771C:p.L924P, NM_178006:exon13:c.T3125C:p.L1042P, NM_178007:exon13:c.T3101C:p.L1034P, NM_001243476:exon17:c.T3020C:p.L1007P)	nonsynonymous
De novo	14:70238173	A	G	SRSF5 (NM_001039465:exon8:c.A814G:p.N272D, NM_006925:exon8:c.A814G:p.N272D)	nonsynonymous
De novo	16:67995561	C	T	SLC12A4 (NM_001145962:exon2:c.G265A:p.V89I, NM_001145961:exon3:c.G259A:p.V87I, NM_001145963:exon3:c.G241A:p.V81I, NM_001145964:exon3:c.G166A:p.V56I, NM_005072:exon3:c.G259A:p.V87I)	nonsynonymous
De novo	16:77353790	G	A	ADAMTS18 (NM_199355:exon16:c.C2488T:p.R830C)	nonsynonymous
Autosomal recessive	17:48356260	G	A	TMEM92 (NM_153229:exon4:c.G269A:p.S90N, NM_001168215:exon5:c.G269A:p.S90N)	nonsynonymous
De novo	19:56284457	G	A	RFPL4AL1(NM_001277397:exon2:c.G776A:p.G259E)	nonsynonymous
De novo	22:22843455	C	A	ZNF280B (NM_080764:exon4:c.G269T:p.S90I)	nonsynonymous
X-linked	X:70621541	T	C	TAF1 (NM_001286074:exon25:c.T4010C:p.I1337T, NM_004606:exon25:c.T4010C:p.I1337T, NM_138923:exon25:c.T3947C:p.I1316T)	nonsynonymous
Comp het	14:64449355	T	C	SYNE2 (NM_015180:exon17:c.T1844C:p.F615S,	nonsynonymous

				NM_182914:exon17:c.T1844C:p.F615S)	
Comp het	14:64557734	A	G	SYNE2 (NM_015180:exon60:c.A11944G:p.N3982D, NM_182914:exon60:c.A11944G:p.N3982D)	nonsynonymous
	14:64557734	A	G	SYNE2 (NM_015180:exon60:c.A11944G:p.N3982D, NM_182914:exon60:c.A11944G:p.N3982D)	nonsynonymous
	14:64608748	A	G	SYNE2 (NM_015180:exon82:c.A15248G:p.D5083G, NM_182914:exon82:c.A15248G:p.D5083G)	nonsynonymous

In summary, SeqHBase is a reliable big data-based computational toolset for efficiently manipulating genome-wide variants, annotations, and every-site coverage in NGS studies. It uses a heuristic framework of inheritance information for detecting *de novo*, inherited homozygous or compound heterozygous mutations that may be disease contributing in trios, nuclear, and/or extended families. It shows very good performance on three different examples of family-based sequencing data, and it is scalable by virtue of its basis on MapReduce framework. SeqHBase is freely available for use by academic or non-profit organizations at http://seqhbase.omicspace.org/. More detailed and updated documents are available on the website.

REFERENCES

Abecasis, G. R. et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491, 56-65.

Adzhubei, I. A. et al. (2010) A method and server for predicting damaging missense mutations. *Nat Methods,* 7, 248-249.

Apache Foundation. Hadoop. http://hadoop.apache.org/. Accessed May 2014.

Apache Foundation. HBase. http://hbase.apache.org/. Accessed May 2014.

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) Accessed May 2014.

Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).

Kumar, P. et al. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc,* 4, 1073-1081.

Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980-985, doi:10.1093/nar/gkt1113 (2014).

Li, H. et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25, 2078-2079.

Wang, K. et al. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res,* 38, e164.