KGGSeq: A biological Knowledge-based mining platform for Genomic and Genetic studies using Sequence data

(Version 0.2)

Miaoxin Li & Hongsheng Gui

Contents

KGGSeq: A biological Knowledge-based mining platform for Genomic and Genetic studies using Sequence data 1
1. Overview 3
1.1 Pipeline for prioritizing genetic variants in rare Mendelian disorders 3
1.2 Pipeline for prioritizing genetic variants in complex diseases 4
1.3 References 5
2. Installation 6
2.1 System requirement 6
2.2 Installing KGGSeq 6
2.3 Installing KGGSeq.Commands.Generator 7
3. Tutorial 8
Example 1: Prioritize variants based on the hg18 assembly for a rare Mendelian disease 8
Example 2: Prioritize variants based on the hg19 assembly for a rare Mendelian disease 10
Example 3: Prioritize variants based on the hg18 assembly for a complex disease 12
4. Reference 14
4.1 Data input and quality control 14
4.3. Prioritize variants for rare Mendelian disorders 17
4.4. Prioritize variants for complex disorders 21
4.5. Output setting and annotation 22
Table 4.5.1 Annotation for list of variables output for KGGSeq pipeline 1 23

Table 4.5.2 Annotation for list of variables output for KGGSeq pipeline 2 26
5. FAQs 28

1. Overview

KGGSeq (http://statgenpro.psychiatry.hku.hk/limx/KGGSeq/) is designed to make use of biological knowledge for sequencing-based gene mapping of human diseases/traits. Compared with other tools like Plink/Seq ("A library for the analysis of genetic variation data", http://atgu.mgh.harvard.edu/plinkseq/), KGGSeq focuses more on downstream analysis (see Figure 1). Currently it mainly has two pipeline carefully designed to filter and prioritize gene variants in exome sequencing of rare Mendelian disorders and common complex disorders. More pipelines will be developed on the KGGSeq platform for the knowledge-based downstream analysis of monogenic and complex diseases/traits using sequencing data.

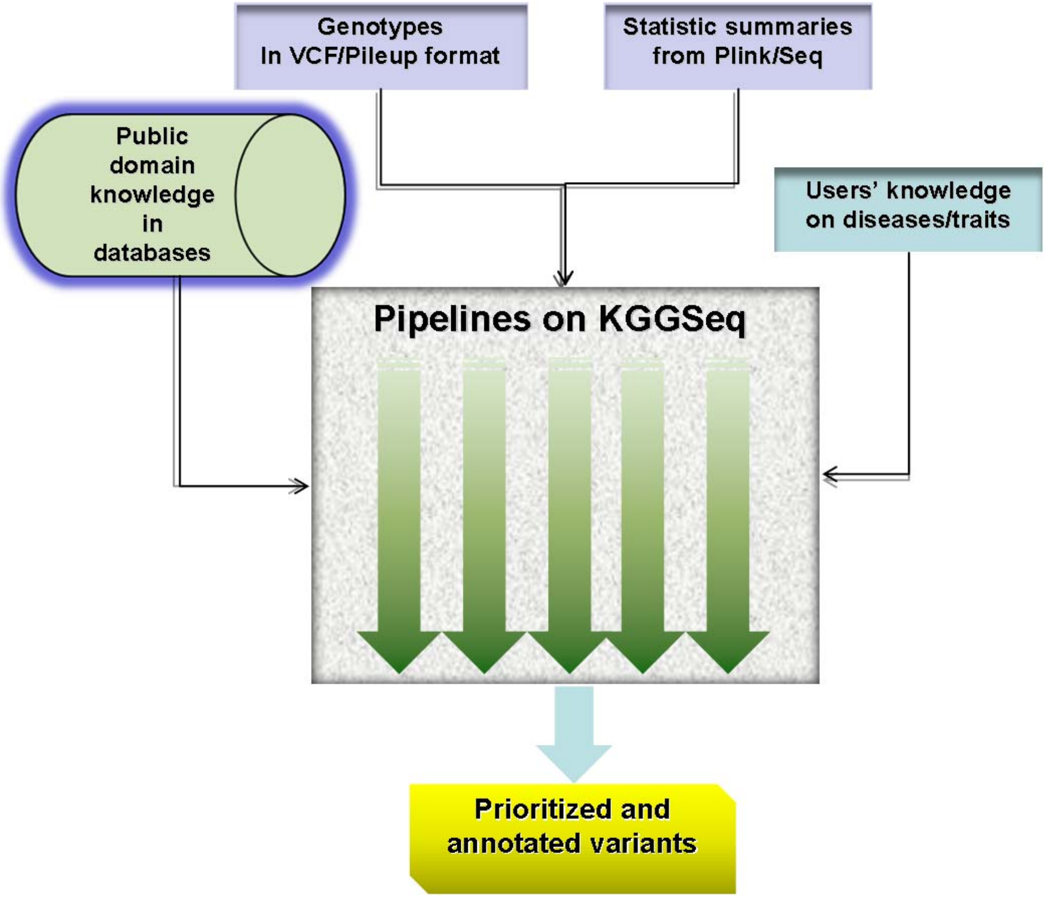


Figure 1. General workflow in KGGSeq

1.1 Pipeline for prioritizing genetic variants in rare Mendelian disorders

A 3-level filtration and prioritization procedure is implemented (see Figure 2).

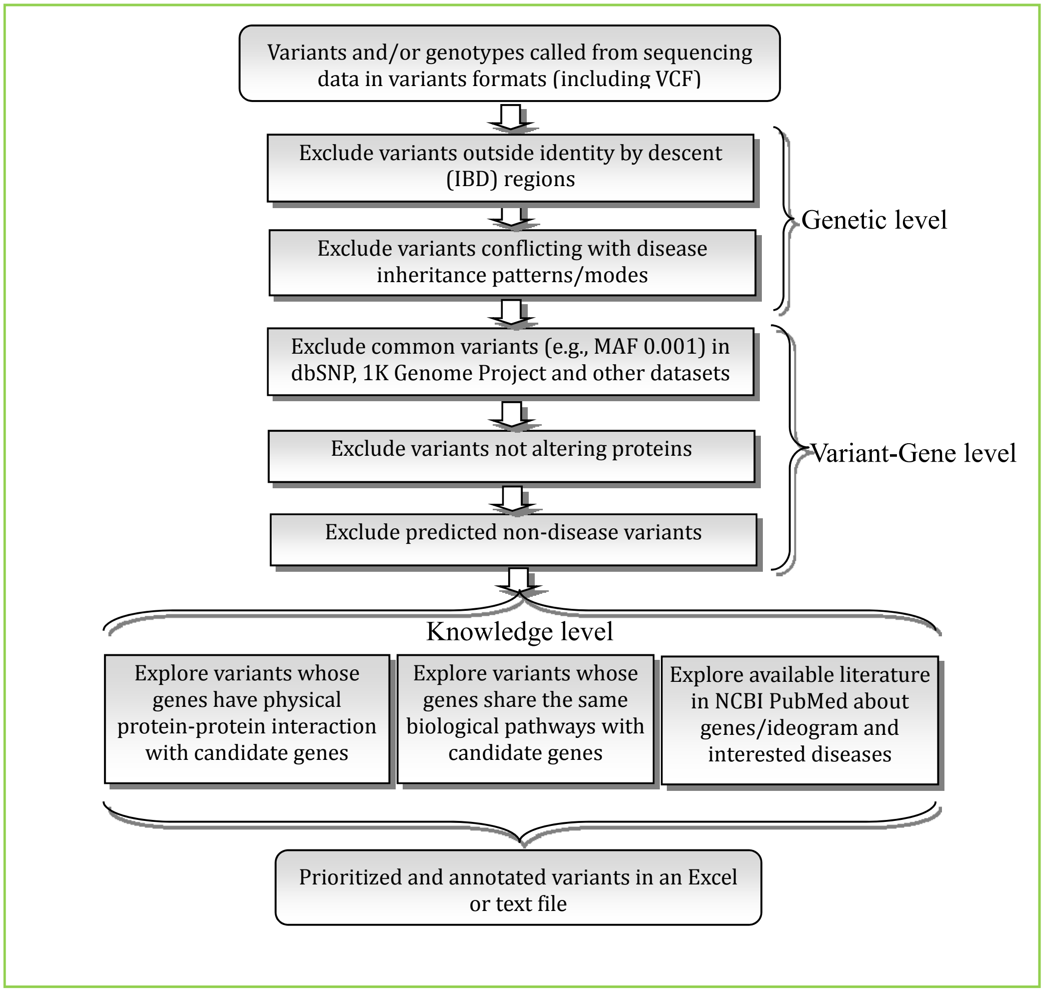

Variants and/or genotypes called from sequencing data in variants formats (including VCF)
Exclude variants outside identity by descent (IBD) regions
Exclude variants conflicting with disease inheritance patterns/modes
Genetic level
Exclude common variants (e.g., MAF 0.001) in dbSNP, 1K Genome Project and other datasets
Exclude variants not altering proteins
Exclude predicted non-disease variants
Variant-Gene level
Knowledge level
Explore variants whose genes have physical protein-protein interaction with candidate genes
Explore variants whose genes share the same biological pathways with candidate genes
Explore available literature in NCBI PubMed about genes/ideogram and interested diseases
Prioritized and annotated variants in an Excel or text file

Figure 2. Pipeline for prioritizing genetic variants in rare Mendelian disorders

The proposed procedure above is comprised of a series of functions to filter and prioritize variants at three classified levels, genetic level, variant-gene level and knowledge level, according to the resources used. Rationales and evaluations of each function are described in the reference paper 1 which has been submitted. We will disclose them once the paper is accepted (sorry about this). On KGGSeq, these functions can be carried out either sequentially or optionally according to various purposes.

1.2 Pipeline for prioritizing genetic variants in complex diseases

A modified pipeline of above 3-level filtration and prioritization procedure is implemented for handling complex diseases (see Figure 3).

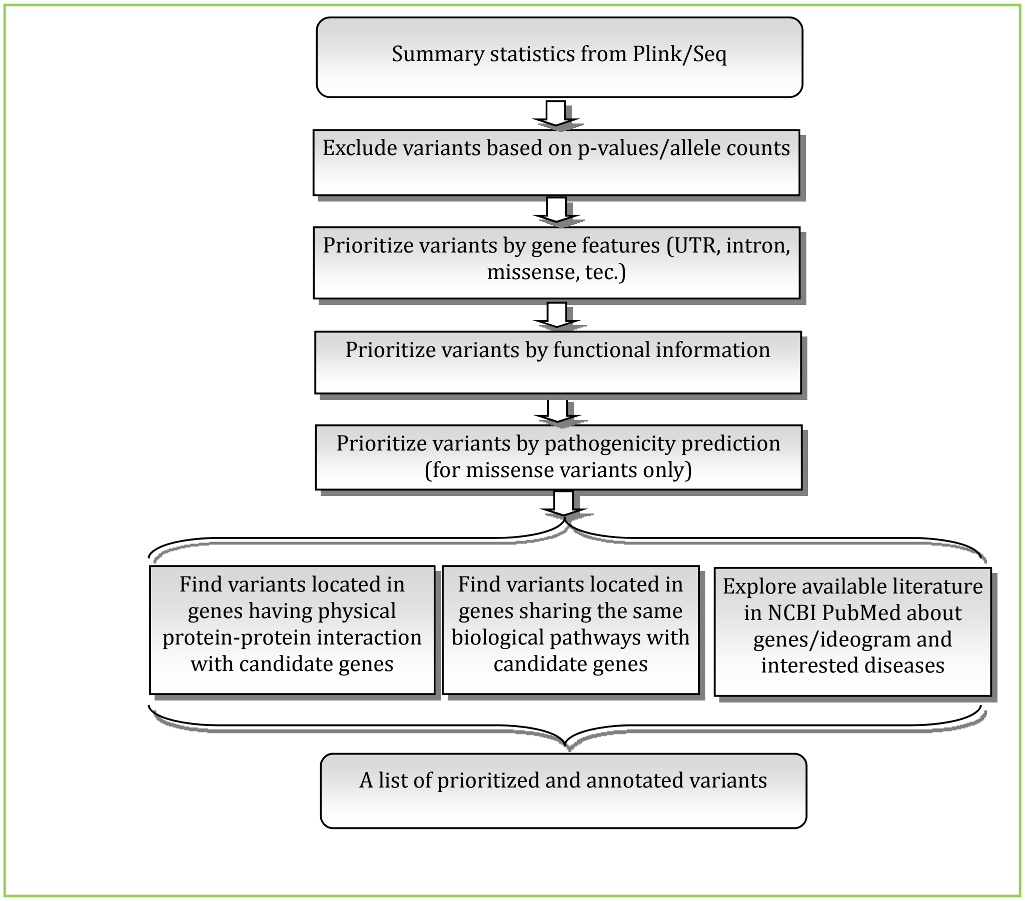

Summary statistics from Plink/Seq
Exclude variants based on p-values/allele counts
Prioritize variants by gene features (UTR, intron, missense, tec.)
Prioritize variants by functional information
Prioritize variants by pathogenicity prediction (for missense variants only)
Find variants located in genes having physical protein-protein interaction with candidate genes
Find variants located in genes sharing the same biological pathways with candidate genes
Explore available literature in NCBI PubMed about genes/ideogram and interested diseases
A list of prioritized and annotated variants

Figure 3. Pipeline for prioritizing genetic variants in complex diseases

Rationales and evaluations of each function are described in the reference paper 1 and 2 which have been submitted. We will disclose them once the paper is accepted (sorry about this). On KGGSeq, these functions can be carried out either sequentially or optionally according to various purposes.

1.3 References

1. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res. 2012 Jan 12. [Epub ahead of print]
2. Li et al. The amount of common and rare missense alleles involved in human diseases are not small. (Submitted)

2. Installation

2.1 System requirement

Java Runtime Environment (JRE) version 6.0 is required for KGGSeq. It can be downloaded from the Java web site.

The version number for KGGSEQ is 1.6 or up. Installing the JRE is very easy in Windows OS and Mac OS X.

In Linux, you have more work to do. Details of the installation can be found http://www.java.com/en/download/help/linux_install.xml.
In Ubuntu, if you have an error message like: “Exception in thread "AWT-EventQueue-0" java.awt.HeadlessException …” , please install the Sun Java Running Environment (JRE) first.

To install the Sun JRE on Ubuntu(10.04), please use the following commands:

```
sudo add-apt-repository “deb http://archive.canonical.com/ lucid partner”

sudo apt-get update

sudo apt-get install sun-java6-jre sun-java6-plugin sun-java6-fonts
```

Detailed explanation of above commands can be found at http://www.ubuntugeek.com/how-install-sun-java-runtime-environment-jre-in-ubuntu-10-04-lucid-lynx.html.

For Mac OS, the JRE 1.6 has been available at http://developer.apple.com/java/download/ since April 2008. Mac OS users may need update the Java application to run IGG. A potential problem is that this update does not replace the existing installation of J2SE 5.0 or change the default version of Java. Similar to the Linux OS, the Java_Home environmental variable has to be configured to initiate KGG.

2.2 Installing KGGSeq

Simply decompress the archive and run the following command

```
java -Xms256m -Xmx1300m -jar "./kggseq.jar” <arguments >
```

The arguments –Xms256m and –Xmx1300m set the initial and maximum Java heap sizes for KGGSeq as 256 megabytes and 1.3 gigabytes respectively. Specifying a larger maximum heap size can speed up the analysis. A higher setting like –Xmx2g is required when there is a large number of variants, say 5 million. The number, however, should be less than the size of physical memory of a machine.

Note: *<arguments >* **can be saved in a flat text file**

2.3 Installing KGGSeq.Commands.Generator

Simply download a file named kggseq.commands.generator (http://statgenpro.psychiatry.hku.hk/limx/kggseq/download/kggseq.commands.generator.jar) the archive and run the following command

```
java -jar "./ kggseq.commands.generator.jar”
```

Note: kggseq.commands.generator.jar is an independent utility with friendly GRAPHIC interface to help users preparing input arguments of KGGSeq.

Note: To use KGGSeq and KGGSeq.Commands.Generator under Proxy Network, you need specifically use the following java command to configure the proxy settings (Thank Daniele Yumi for this suggestion):

```
java -Dhttp.proxyHost=xxx.xxx.xxx -Dhttp.proxyPort=xxxx -Dhttp.proxyUser=xxx
-Dhttp.proxyPassword=xxx -jar “./kggseq.commands.generator.jar”
```

3. Tutorial

Example 1: Prioritize variants based on the hg18 assembly for a rare Mendelian disease

Files needed:

1) a VCF file (rare.disease.hg18.vcf); and
2) a linkage pedigree file (rare.disease.ped.txt).

Note: All files were included in the examples folder of KGGSeq.

Run the command below:

```
java -jar -Xms256m -Xmx1300m kggseq.jar examples/param.rare.disease.hg18.txt
```

We now walk through the parameter file "param.rare.disease.hg18.txt" before going into the results. Lines starting with hash sign # are comments. Detailed interpretation for each argument in the parameter file is included in Chapter 4.

```
#one argument per line

#Choose KGGSeq pipeline type
--prioritize-monogen \ #line 1

#Specify the input files
--vcf-file examples/rare.disease.hg18.vcf \ #line 2
--ped-file examples/rare.disease.ped.txt \ # line 3, or specify '--indiv-pheno
C:1,B:1,A:2'

#QC
--seq-qual 50 \ #line 4
--gty-qual 10 \ #line 5
--seq-dp 4 \ #line6
--no-missing-case \ #line 7

#general setting
--buildver hg18 \ #line 8

#Genetic level
--genotype-filter 3 \    \ #line 9 this can be multiple numbers like 3,4
--ibs-check \ # line 10, or specify '--ibd-anno ibdregions.txt'
```

```
#Variants level
--db-gene refgene \ # line 11
--gene-feature-in 0,1,2,3,4,5 \ #line 12
--db-filter
hg18_1kgasn2010,hg18_1kgeur2010,hg18_1kgafr2010,hg18_1kgasn2009,hg18_1kgeur2009,hg18_1kgafr2009,hg18_dbsnp130,hg18_dbsnp129 \ # line 13
#--local-filter file1/to/path, file2/to/path
--rare-allele-freq 0.006 \ #line 14
--db-score dbnsfp \ #line 15
--filter-nondisease-variant \ #line 16

#Literature level
--candi-list ATXN1,ATXN2,ATXN8OS,ATXN8,ATXN10,TTBK2 \ #line 17
--pathway cura \ or cano # cura and cano indicate different pathway datasets of GSEA
--ppi string \    # string indicates the STRING PPI database
--pubmed-mesh Spinocerebellar+ataxia \ #line 18

#Gene level
--pseudogene-anno \ #line 19
--dispengene-anno \ #line 20

#Output setting
--out ./test1 \ #line 21
--excel \ # line 22
```

Part (I): choose KGGSeq pipeline type

Argument ‘line 1’ is used to select the KGGSeq pipeline 1 (see chapter 1) for analyzing variants for rare mendelian diseases.

Part (II): specify the input files and adopt QC on raw variants

Arguments ‘line 2~7’ are used to specify the input file (VCF format) and adopt basic quality control on the raw variants.

Part (III): general setting

Argument ‘line 8’ is used to set the human genome build version at hg18, which should be consistent with the upstream analysis for the raw variants (mapping and calling).

Part (IV): genetic level filtration

Arguments ‘line 9~10’ are used to apply ‘genetic level filtration’ on the variants, including genetic model setting and IBS estimation.

Part (V): variant level filtration

Arguments ‘line 11~16’ are used to apply ‘variant level filtration’ on the variants, including removal of variants in non-exonic regions, covered by common databases with relative high frequency and predicted as ‘non-disease causal’.

Part (VI): literature level annotation

Arguments ‘line 17~18’ are used to annotate remained list of variants/genes by searching shared protein-protein interaction or pathway with provided candidate genes, and previous literature evidence.

Part (VII): gene level annotation

Arguments 'line 19~20' are used to annotate remained list of variants/genes by searching overlapping pseudogenes or dispensable genes (compiled from 1000 genome project).

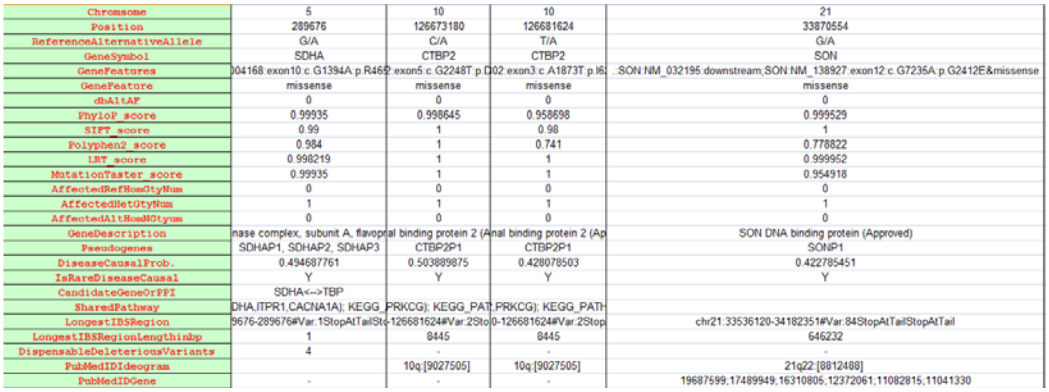
Part (VIII): output setting

Arguments 'line 21~22' are used to set the path, format and name of KGGSeq output file.

☺ Tips: most of the above arguments are optional (except line 1, 2 and 3), so user can mask some lines by "#" or delete the lines. Under this circumstance, user can have a systematic view of the impact for each level or even steps. And it will be easier to produce this parameter file by Kggseq command generator we provide.

After running KGGSeq, you will get one file named '**test1.fit.xls**' in the same directory of KGGSeq.jar. 4 missense variants residing in 3 genes (SDHA, CTBP2 and SON) are retained and all of them are predicted to be disease causal. The results are also supported by literature and our knowledge of PPI and pathway (see table 3.1).

Table 3.1 Output results for example 1



Chromsome	5	10	10	21
Position	289676	126673180	126681624	33870554
ReferenceAlternativeAllele	G/A	C/A	T/A	G/A
GeneSymbol	SDHA	CTBP2	CTBP2	SON
GeneFeatures	004168:exon10:c.G1394A:p.R46	2:exon5:c.G2248T:p.D	02:exon3:c.A1873T:p.I6	..:SON:NM_032195:downstream;SON:NM_138927:exon12:c.G7235A:p.G2412E&missense
GeneFeature	missense	missense	missense	missense
dbAltAF	0	0	0	0
PhyloP_score	0.99935	0.998645	0.958698	0.999529
SIFT_score	0.99	1	0.98	1
Polyphen2_score	0.984	1	0.741	0.778822
LRT_score	0.998219	1	1	0.999952
MutationTaster_score	0.99935	1	1	0.954918
AffectedRefHomGtyNum	0	0	0	0
AffectedHetGtyNum	1	1	1	1
AffectedAltHomNGtyum	0	0	0	0
GeneDescription	nase complex, subunit A, flavopr	al binding protein 2 (A	nal binding protein 2 (Ap	SON DNA binding protein (Approved)
Pseudogenes	SDHAP1, SDHAP2, SDHAP3	CTBP2P1	CTBP2P1	SONP1
DiseaseCausalProb.	0.494687761	0.503889875	0.428078503	0.422785451
IsRareDiseaseCausal	Y	Y	Y	Y
CandidateGeneOrPPI	SDHA<-->TBP			
SharedPathway	DHA,ITPR1,CACNA1A); KEGG_	PRKCG); KEGG_PAT	PRKCG); KEGG_PATH	
LongestIBSRegion	9676-289676#Var:1StopAtTailSt	-126681624#Var:2Sto	0-126681624#Var:2Stop	chr21:33536120-34182351#Var:84StopAtTailStopAtTail
LongestIBSRegionLengthinbp	1	8445	8445	646232
DispensableDeleteriousVariants	4	-	-	-
PubMedIDIdeogram		10q:[9027505]	10q:[9027505]	21q22:[8812488]
PubMedIDGene	-	-	-	19687599;17489949;16310805;12372061;11082815;11041330

Notes: 1) the original excel sheet was transposed; 2) detailed annotation for each 'item' in the first column are listed in Table 4.5.1 of chapter 4.

Example 2: Prioritize variants based on the hg19 assembly for a rare Mendelian disease

Files needed:

1) a VCF file (rare.disease.hg19.vcf); and
2) a pedigree file (rare.disease.ped.txt).

Note: All files were included in the example folder of KGGSeq.

Run the command below:

```
java -jar -Xms256m -Xmx1300m kggseq.jar examples/param.rare.disease.hg19.txt
```

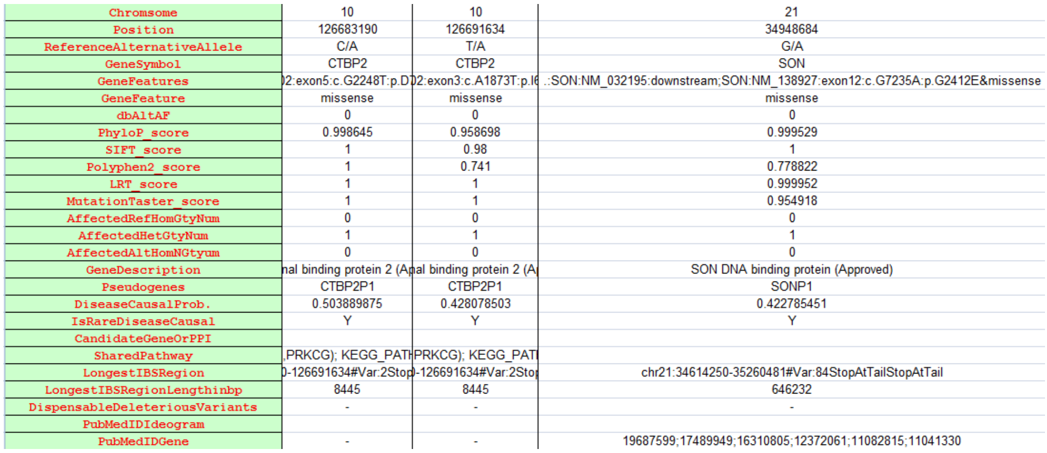
We now walk through the parameter file “param.rare.disease.hg19.txt” before going into the results. Lines starting with hash sign # are comments. Detailed interpretation for each argument in the parameter file is included in chapter 4.

```
#one argument per line

#Choose KGGSeq pipeline type
--prioritize-monogen \

#specify the input files
--vcf-file rare.disease.hg19.vcf \
--ped-file rare.disease.ped.txt # or --indiv-pheno C:1,B:1,A:2 \

#QC
--seq-qual 50 \
--gty-qual 10 \
--seq-dp 4 \
--no-missing-case \

#General setting
--buildver hg19 \

#Genetic level
--genotype-filter 3 \ this can be multiple numbers like 3,4
--ibs-check \ # or specify --ibd-anno ibdregions.txt

#Variants level
--db-gene refgene \
--gene-feature-in 0,1,2,3,4,5 \
--db-filter hg19_1kg201011,hg19_1kg201105,hg19_dbsnp131 \
--rare-allele-freq 0.006 \
--db-score dbnsfp \
--filter-nondisease-variant \

#Literature level
--candi-list ATXN1,ATXN2,ATXN3,SCA4,SCA16,TBP \
--pubmed-mesh Spinocerebellar+ataxia \

#Gene level
--pseudogene-anno \
--dispengene-anno \

#Output setting
--out ./test2 \
--excel
```

The major difference of example 2 from example 1 is to set human genome build version at hg19.

After running KGGSeq for example 2, one file named ‘test2.fit.xls’ will be saved in the current directory. Three variants residing in two genes (CTBP2 and SON) are remained and well annotated (see table 3.2).

Table 3.2 Output results for example 2



Chromsome	10	10	21
Position	126683190	126691634	34948684
ReferenceAlternativeAllele	C/A	T/A	G/A
GeneSymbol	CTBP2	CTBP2	SON
GeneFeatures	)2:exon5:c.G2248T:p.D	)2:exon3:c.A1873T:p.I6	.:SON:NM_032195:downstream;SON:NM_138927:exon12:c.G7235A:p.G2412E&missense
GeneFeature	missense	missense	missense
dbAltAF	0	0	0
PhyloP_score	0.998645	0.958698	0.999529
SIFT_score	1	0.98	1
Polyphen2_score	1	0.741	0.778822
LRT_score	1	1	0.999952
MutationTaster_score	1	1	0.954918
AffectedRefHomGtyNum	0	0	0
AffectedHetGtyNum	1	1	1
AffectedAltHomNGtyum	0	0	0
GeneDescription	nal binding protein 2 (A	al binding protein 2 (A	SON DNA binding protein (Approved)
Pseudogenes	CTBP2P1	CTBP2P1	SONP1
DiseaseCausalProb.	0.503889875	0.428078503	0.422785451
IsRareDiseaseCausal	Y	Y	Y
CandidateGeneOrPPI			
SharedPathway	,PRKCG); KEGG_PATH	PRKCG); KEGG_PATI	
LongestIBSRegion	0-126691634#Var:2Stop	-126691634#Var:2Stop	chr21:34614250-35260481#Var:84StopAtTailStopAtTail
LongestIBSRegionLengthinbp	8445	8445	646232
DispensableDeleteriousVariants	-	-	-
PubMedIDIdeogram			
PubMedIDGene	-	-	19687599;17489949;16310805;12372061;11082815;11041330

Notes: 1) the original excel sheet was transposed; 2) detailed annotation for each ‘item’ in the first column are listed in Table 4.5.1 of chapter 4.

Example 3: Prioritize variants based on the hg18 assembly for a complex disease

Files needed:

1) a Plink/Seq summary result file (complex.disease.v.assoc.hg18.out)

Note: All files were included in the example folder of KGGSeq.

Run the command below:

```
java -jar -Xmx1300m kggseq.jar examples/param.complex.disease.v.assoc.hg19.txt
```

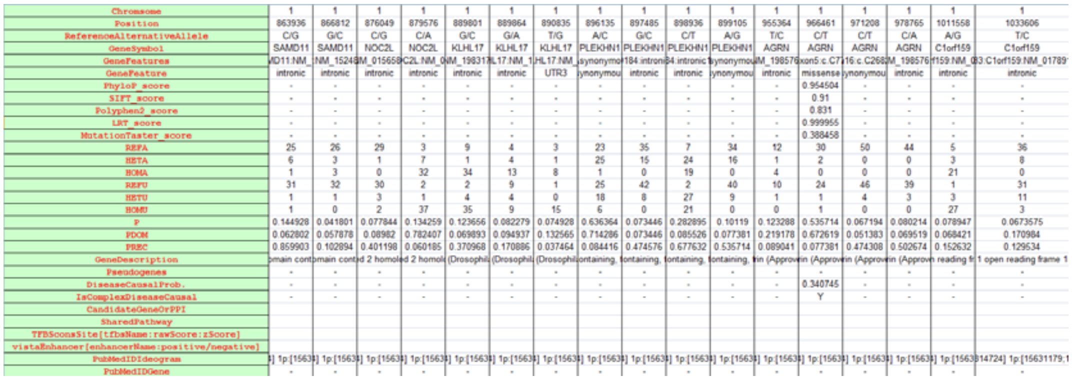
We now walk through the parameter file “param.complex.disease.v.assoc.hg19.txt” before going into the results. Lines starting with hash sign # are comments. Detailed interpretation for each argument in the parameter file is included in chapter 4.

```
#Choose KGGSeq pipeline type
--prioritize-complex

#specify the input file
--v-assoc-file    complex.disease.v.assoc.hg18.out

#QC
```

```
--min-obsa 10
--min-obsu 10

#filtration by summary results
--filter-model association    # case-unique, control-unique, association or no
--p-value-cutoff 0.1 #only effective for 'association' filter-model
--qqplot #only effective for 'association' filter-model

#filtration by knowledge
--db-gene refgene \
--gene-feature-in 0,1,2,3,4,5,6,7,8,9,10

--db-score dbnsfp
--enhancer-anno
--tfbs-anno

--candi-list HLA-DPA1,HLA-DPB1
--pathway cura \ or cano # cura and cano indicate different pathway datasets of GSEA
--ppi string \    # string indicates the STRING PPI database
--pubmed-mesh Crohn's+disease,Inflammatory+bowel+disease

#output setting
--buildver hg18
--excel
--out ./test3 \
```

After running KGGSeq for example 3, one file named ‘test3.fit.xls’ and one figure named ‘test3.qq.png’ will be saved in the current directory. In the excel file, 17 variants (only 1 missense variant) residing in 6 genes (SAMD11, NOC2L, KLHL17, PLEKHN1, AGRN and C1orf159) are remained and well annotated (see table 3.3).

Table 3.2 Output results for example 3



Chromsome	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Position	863936	866812	876049	879576	889801	889864	890835	896135	897485	898936	899105	955364	966461	971208	978765	1011558	1033606
ReferenceAlternativeAllele	C/G	G/C	C/G	C/A	G/C	G/A	T/G	A/C	G/C	C/T	A/G	T/C	C/T	C/T	C/A	A/G	T/C
GeneSymbol	SAMD11	SAMD11	NOC2L	NOC2L	KLHL17	KLHL17	KLHL17	PLEKHN1	PLEKHN1	PLEKHN1	PLEKHN1	AGRN	AGRN	AGRN	AGRN	C1orf159	C1orf159
GeneFeatures	MD11:NM_	NM_15248	M_015658	C2L:NM_0	M_198317	L17:NM_1	HL17:NM_	synonymo	184:intron	84:intronic	synonymou	M_198576	xon5:c.C7	16:c.C268	M_198576	f159:NM_0	3:C1orf159:NM_01789
GeneFeature	intronic	intronic	intronic	intronic	intronic	intronic	UTR3	ynonymou	intronic	intronic	ynonymou	intronic	missense	ynonymou	intronic	intronic	intronic
PhyloP_score	-	-	-	-	-	-	-	-	-	-	-	-	0.954504	-	-	-	-
SIFT_score	-	-	-	-	-	-	-	-	-	-	-	-	0.91	-	-	-	-
Polyphen2_score	-	-	-	-	-	-	-	-	-	-	-	-	0.831	-	-	-	-
LRT_score	-	-	-	-	-	-	-	-	-	-	-	-	0.999955	-	-	-	-
MutationTaster_score	-	-	-	-	-	-	-	-	-	-	-	-	0.388458	-	-	-	-
REFA	25	26	29	3	9	4	3	23	35	7	34	12	30	50	44	5	36
HETA	6	3	1	7	1	4	1	25	15	24	16	1	2	0	0	3	8
HOMA	1	3	0	32	34	13	8	1	0	19	0	4	0	0	0	21	0
REFU	31	32	30	2	2	9	1	25	42	2	40	10	24	46	39	1	31
HETU	1	1	3	1	4	4	0	18	8	27	9	1	1	4	3	3	11
HOMU	1	0	2	37	35	9	15	6	0	21	0	0	1	0	0	27	3
P	0.144928	0.041801	0.077844	0.134259	0.123656	0.082279	0.074928	0.636364	0.073446	0.282895	0.10119	0.123288	0.535714	0.067194	0.080214	0.078947	0.0673575
PDOM	0.062802	0.057878	0.08982	0.782407	0.069893	0.094937	0.132565	0.714286	0.073446	0.085526	0.077381	0.219178	0.672619	0.051383	0.069519	0.068421	0.170984
PREC	0.859903	0.102894	0.401198	0.060185	0.370968	0.170886	0.037464	0.084416	0.474576	0.677632	0.535714	0.089041	0.077381	0.474308	0.502674	0.152632	0.129534
GeneDescription	omain cont	omain cont	d 2 homolo	d 2 homolo	(Drosophil	(Drosophil	(Drosophil	ontaining,	ontaining,	ontaining,	ontaining,	in (Approv	in (Approv	in (Approv	in (Approv	n reading fr	1 open reading frame 1
Pseudogenes	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DiseaseCausalProb.	-	-	-	-	-	-	-	-	-	-	-	-	0.340745	-	-	-	-
IsComplexDiseaseCausal	-	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	-	-
CandidateGeneOrPPI																	
SharedPathway																	
TFBSconsSite[tfbsName:rawScore:zScore]																	
vistaEnhancer[enhancerName:positive/negative]																	
PubMedIDIdeogram	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[1563	1p:[15631179;1
PubMedIDGene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Notes: 1) the original excel sheet was transposed; 2) detailed annotation for each ‘item’ in the first column are listed in Table 4.5.2 of chapter 4.

4. Reference

4.1 Data input and quality control

For variants and genotypes inputted in VCF format

ID	Arguments/values	Annotation
1	--vcf-file /path/to/file	The genotype dataset with vcf format; it can also be compressed as .gz or .zip. [compulsory] *NOTE: If the data were stored in different files chromosome by chromosome, you can use _CHROM_ to denote the chromosome names [1…Y] or directly specify* [1,2,X]*in the file name so that kggseq can read data in multiple files in a run.*
2	--indiv-pheno indivID1:0,indivID2:2,…	Specify the individual IDs (which are identical to those in the vcf file) and affection status, by default, which are coded: 0 missing or unknow,1 unaffected,2 affected. [either option 2 or 3 is compulsory]
3	--ped-file path/to/pedigree/file	Specify the individual IDs and relationship between subjects; it is a conventional linkage format (http://www.broadinstitute.org/science/programs/medical-and-population-genetics/haploview/input-file-formats-0). [either option 2 or 3 is compulsory]
4	--seq-qual 50	Set minimal average base call quality (PhredQualityScore) of each variant; it is 50 by default. --seq-qual 50 means QUAL>= 50 for further processing. [optional]
5	--gty-qual 10	Set minimal average genotyping quality (PhredQualityScore) of each variant; it is 10 by default. --gty-qual 10 means GQ >= 10 for further processing. [optional]
6	--gty-dp 4	Set minimal sequence depth at a site; it is 4 by default. --seq-dp 4 means DP (of a genotype)>= 4 for further processing [optional]
7	--gty-af-ref 0.05	Set the maximal fractions of reads supporting each reported alternative allele in reference-allele homozygous genotypes; it is 0.05 by default. --gty-af-ref 0.05 means AF (of a genotype)<= 0.05 or the a proportion of alternative alleles estimated by the AD)<= 0.05 at a reference-allele homozygous genotype for further processing [optional] Note: The AF is prioritized in the quality control process if the both AF and Ad are provided in the VCF file as the AF is usually estimated using high quality reads.
8	--gty-af-alt 0.25	Set the minimal fractions of reads supporting each reported alternative allele in all non-reference-allele homozygous or heterozygous genotypes; it is 0.25 by default. --gty-af-alt 0.25 means AF (of a genotype)>= 0.25 or the a proportion of alternative alleles estimated by the AD)>= 0.25 at a

		non-reference-allele homozygous or heterozygous genotype for further processing [optional] Note: The AF is prioritized in the quality control process if the both AF and Ad are provided in the VCF file as the AF is usually estimated using high quality reads.
9	--disable-vcf-filter	Ignore the FILTER tag in CVF file [optional]
10	--no-missing-case	Remove those variants having missing data in cases. [optional]
11	--regions-in chr2,chr4:21212-233454,chrX	Specify the interested chromosomes or regions to be included
12	--regions-out chr2,chr4:21212-233454,chrX	Specify the chromosomes or regions to be excluded
	--ignore-indel	Ignore the Insertion or deletion variants in a VCF file
	--ignore-snv	Ignore the single nucleotide variants in a VCF file

Notes: SoapSNP, ANNOVAR and Samtools Pileup data can be inputted by setting '--soap-file', '--annovar-file' and '--pileup-file' instead of '--vcf-file'.

☺ *Tips: VCF format, SoapSNP format, Samtools pileup format and ANNOVAR format are previously described; user can check the weblinks (http://vcftools.sourceforge.net/specs.html, http://soap.genomics.org.cn/soapsnp.html, http://samtools.sourceforge.net/pileup.shtml, and http://www.openbioinformatics.org/annovar/annovar_input.html) for detail. And if the Soap SNP data are stored in different files according to chromosomes and individual IDs, you can use _CHROM_ and _INDIVID_ to denote chromosome names and individual names in the file or folder names so that kggseq can read genotypes in multiple files in a run.*

For variants/genes inputted in summary results of plink/seq

ID	Arguments/values	Annotation
1	--v-assoc-file /path/to/file	The output of plink/seq v-assoc in a text file; it can also be compressed as .gz or .zip. [compulsory]
2	--min-heta 1	Set minimal observed number of heterozygotes in cases (the affected). [optional]
3	--min-homa 1	Set minimal observed number of alternate homozygsote(s) in cases (the affected); option 2 and 3 are logical OR relation. [optional]
4	--min-hetu 1	Set minimal observed number of heterozygotes in controls (the unaffected). [optional]
5	--min-homu 1	Set minimal observed number of alternate

		homozygsote(s) in controls(the unaffected; Arg.4 and 5 are logical OR relation. [optional]
6	--min-obsa 2	Set minimal observed number of non-null genotypes in cases (the affected). [optional]
7	--min-obsu 2	Set minimal observed number of non-null genotypes in controls(the unaffected); Arg.6 and 7 are logical AND relation. [optional]
8	--min-obs 2	Set minimal observed number of non-null genotypes in all samples. [optional]
9	--assoc-file /path/to/file	The output of plink/seq of gene-based association analysis result in a text file; it can also be compressed as .gz or .zip. [compulsory] Note: For this type of file arguments of 2 to 8 are not necessary

☺*Tips: Procedures for running plink/seq to get the summary statistics:*

Prepare input by plink/seq

1.1 For individual variants

pseq proj v-assoc --phenotype my.phenotype *[and other filter options]*

OR

pseq data.vcf.gz v-assoc --phenotype phenofile.txt my.phenotype *[and other filter options]*

1.2 For groups of variants or genes

pseq proj assoc --phenotype my.phenotype --options calpha vt --mask loc.group=refseq *[other filter options]*

For genes inputted in summary results of VAAST software tool (http://www.yandell-lab.org/software/vaast.html)

ID	Arguments/values	Annotation
1	--vaast-simple-file /path/to/file	The output of VAAST of gene-based association analysis result in a text file; it can also be compressed as .gz or .zip. [compulsory] The following are the example format: *RANK Gene p-valueScore Variants* *1 TEKT4 2.84070221452526e-12 47.62418618 chr2:94906054;47.62;T->A;M->K;0,8* *2 HLA-C_DUP_02 2.84070221452526e-12 42.05922817 chr6:31345752;42.06;T->C;T->A;0,12* *3 USP6 2.84070221452526e-12 41.80136081 chr17:4977987;41.80;G->A;V->I;0,8*

4.2 General settings

ID	Arguments	Annotation

1	--buildver hg18	Set the coordinate version of the reference genome; it is hg18 by default. KGGSeq can only support hg18 or higher version. [optional]
2	--resource path/to/resources	Set the path of resource data. [optional]
3	--no-lib-check	Disable library checking at each initiation. [optional]
4	--no-resource-check	Disable resources checking at each initiation. [optional]

4.3. Prioritize variants for rare Mendelian disorders

<table>
<tr><th>ID</th><th>Arguments/values</th><th>Annotation</th></tr>
<tr><td colspan="3"><b>Select pipeline type</b></td></tr>
<tr><td>1</td><td>--prioritize-monogen</td><td>Prioritize variants for monogenic disorders by the three-level filtration and prioritization procedure. But it has to go with the specific settings as described in the following. [compulsory]</td></tr>
<tr><td colspan="2"><b>Set inheritance pattern filter</b></td><td></td></tr>
<tr><td>2</td><td>--genotype-filter 1,2
Note: The inheritance mode based filtration is proposed under strong assumptions for rare Mendelian disease with clear inheritance mode. If the inheritance mode is elusive, such filtration is not suggested as it may lead the miss of genuine mutation(s). This is particularly true when one has no sufficient information to distinguish the <b>compound-heterozygosity</b> diseases from the <b>recessive</b> diseases.</td><td>Exclude variants for which their genotypes are not consistent with the assumption of disease inheritance pattern [optional]
<table>
<tr><th><i>Code</i></th><th><i>Function</i></th><th><i>Applicable model</i></th></tr>
<tr><td><i>1</i></td><td><i>Exclude variants which have heterozygous genotypes in one or more affected family members</i></td><td><i>Recessive</i></td></tr>
<tr><td><i>2</i></td><td><i>Exclude variants which have the same homozygous genotypes in both affected and unaffected family members</i></td><td><i>Recessive and full penetrance causal mutation(s) in the sample</i></td></tr>
<tr><td><i>3</i></td><td><i>Exclude variants which have reference homozygous genotype in one or more affected family members</i></td><td><i>Rare dominant and no consanguineous mating or <b>compound-heterozygosity</b></i></td></tr>
<tr><td><i>4</i></td><td><i>Exclude variants which have the same heterozygous genotypes in both affected and unaffected family members</i></td><td><i>Dominant; full penetrance causal mutation(s) in the sample</i></td></tr>
</table></td></tr>
</table>

5	*Exclude variants which have alternative homozygous genotype in one or more affected family members*	*Rare dominant and no consanguineous mating or* ***compound-heterozygosity***
6	*Exclude variants which have NO shared alleles in affected family members*	*Full penetrance causal mutation(s) in the sample*

IBD estimation		
3	--ibs-check	Check the longest IBS region covering each interesting variant, in which there is at least one allele identical among all cases. For subjects of a pedigree, variants with a long IBS region can be highlighted. [optional] **IBS is over-estimation of IBD. For rare dominant diseases, --ibs-check can be used to guess the shared IDB region.**
4	--ibd-anno ibd/file/path	Read IBD or significant linkage regions derived by third party tools (such as Beagle). The format is: *CHR START END* *1 6134174 10082841* *1 202855724 207862411* Title line is needed. Variants within these regions will be highlighted. [optional]
	--homozygosity-check	Check the longest homozygous region covering each interesting variant. Variants with a long homozygous region can be highlighted. [optional] **For rare dominant diseases, this function can be used to examine the loss of heterozygosity straightforwardly. For rare recessive diseases, it can be used to examine the runs of homozygosity straightforwardly.** But it currently does not consider statistical uncertainty.
Filtration according to allele frequencies		
5	--db-filter hg18_1kgasn2010, hg18_kgeur2010,…	Set databases for filtration. Variants in these databases will be excluded. The database name can be hg18_1kgasn2010,hg18_1kgeur2010,hg18_1kgafr2010, hg18_1kgasn2009,hg18_1kgeur2009,hg18_1kgafr2009, hg18_dbsnp130,hg18_dbsnp129,hg19_1kg201011, hg19_1kg201105,hg19_dbsnp131. No spaces are allowed between and within database names. In addition, we also provided a public variants dataset from NHLBI GO Exome Sequencing Project (ESP, http://evs.gs.washington.edu/EVS/) as reference: hg18_ESP5400 and hg19_ESP5400.snps [optional]

6	--local-filter path/name1,path/name2,...	Set local datasets to filter out variants. Each file has 5 columns [Chromosome, Physical Position, Reference Allele, Alternative Allele(s), and Frequency of alternative allele(s)], separated by tab character. [optional] The format is: *Chrom. Position Reference Alternative Freq* *1 469 C G 0.150* *1 492 C T 0.175* *1 519 G C 0.067* *1 874290 - 0ACAGAG 0.809* *1 875913 CAG 3 0.8431* Title line is needed and Freq is optional. Commands of kggseq to convert the vcf format into the this format: *java -jar kggseq.jar --make-filter --vcf-file path/to/file --buildver-in hg18 --out test.hg18.var --buildver-out hg19* *NOTE: If the data were stored in different files chromosome by chromosome, you can use _CHROM_ to denote the chromosome names [1…Y] or directly specify* [1,2,X]*in the file name so that kggseq can read data in multiple files in a run.*
	--local-filter-vcf path/to/vcffile1, path/to/file2,…	Use in-house VCF data as controls to filter out sequence variants.
7	--rare-allele-freq 0.005	Keep rare variants and filter out common variants (Minor allele frequency over 0.005); it is --rare-allele-freq 0 by default. [optional]
	--allele-freq 0,0.01	Keep the sequence variants with the specified range of MAF (including both boundaries); it is --allele-freq 0,0.01 by default. But this option and --rare-allele-freq are mutually exclusive and the latter has higher priority. [optional]
Filtration according to gene features		
8	--db-gene refgene	Set database(s) to annotate and filter variants. Currently KGGSeq only considers the RefGene database. It can be automatically updated with UCSC annotation database. [optional]
9	--gene-feature-in 0,1,2,3,4,5,…	Variants beyond --gene-feature-in region will be excluded. It is '--gene-feature-in 0,1,2,3,4,5…,15' by default. See Table 4.2 for detail. [optional]
Filtration according to prediction		
10	--db-score dbnsfp	Use an integrated database of functional predictions from multiple algorithms human non-synonymous SNPs databases (dbNSFP, http://sites.google.com/site/jpopgen/dbNSFP) to score variants. These scores will finally be combined by Logistic regression method to predict whether a variant can be disease-casual or not. [optional]
11	--filter-nondisease-variant	Filter out variant predicted to be non-disease causal variant. [optional]

Prioritization according to PPI and pathway sharing		
12	--candi-list GeneSymol1, GeneSymol2,GeneSymol3	Highlight variants of candidate genes or genes having PPI or sharing the same PATHWAYs with these candidate genes in the output. No spaces are allowed between and within gene symbols. [optional]
12’	--candi-file /path/to/file	The same as --candi-list. But specify candidate genes stored in a file. [optional]
	--ppi string	Specify a protein-protein interaction (PPI) database. Currently KGGSeq only considers the STRING PPI database, http://string-db.org/.
	--ppi-depth 2	Specify the depth of the exploration in a PPI network. Here the depth is equal to the number of smallest arcs between two vertices in a PPI network.
	--pathway cura	Specify a pathway dataset. Currently KGGSeq only considers two pathway datasets, C2: curated gene sets(denoted as ‘cura’) and CP: canonical pathways (denoted as ‘cano’), from GSEA. Details of description of the pathways can be found at http://www.broadinstitute.org/gsea/msigdb/collections.jsp#C2
Prioritization according to Literature		
13	--pubmed-mesh Type+1+Diabetes, Type+2+Diabetes,...	Explore the relevant literature in NCBI PubMed about the gene symbol and idogram of a variant and the specified disease names. Preferably the disease name is an official MeSH Term. Please use + to replace all spaces (i.e., blank character) within a disease name term. Note: it will take a long time to retrieve this information of hundreds of variants from NCBI database through internet. [optional]
Prioritization according to Gene information		
14	--pseudogene-anno	Annotate a variant’s gene if it has psedudogene(s) registered at http://tables.pseudogene.org/set.py?id=Human61. This can give users caution of possible mapping error due to these pseudogenes [optional]
15	--dispengene-anno	Annotate a variant’s gene whether it is a possible dispensable gene (compiled from 1000 genome project). If it is, this may indicate the mutation in this gene would not cause any severe Mendelian diseases[optional]
Prioritization according to regulation information		
18	--tfbs-anno	Annotate variants within transfactor binding sites, provided by http://hgdownload.cse.ucsc.edu/downloads.html
19	--filter-nontfbs-variant	Filter out variant beyond ransfactor binding sites
20	--enhancer-anno	Annotate variants within enhancer regions, provided by http://hgdownload.cse.ucsc.edu/downloads.html

21	--filter-nonEnhancer-variant	Filter out variant beyond enhancer regions

Table 4.2 Illustration of different gene features

Feature	Value	Explanation
Frameshift	0	Short insertion or deletion result in a completely different translation from the original.
Nonframeshift	1	Short insertion or deletion result in loss of amino acids in the translated proteins.
Stoploss	2	Indels or nucleotide substitution result in the loss of stop codons (TAG, TAA, TGA)
Stopgain	3	Indels or nucleotide substitution result in the new stop codons (TAG, TAA, TGA), which may truncate the protein.
Missense	4	Variants result in a codon coding for a different amino acid (missense)
Splicing	5	variant is within 2-bp of a splicing junction (use --splicing x to change this, the unit of x is base-pair)
Synonymous	6	Nucleotide substitution does not change amino acid.
Exonic	7	Due to loss of sequences, only map a variant into exonic region
UTR5	8	variant within a 5' untranslated region
UTR3	9	variant within a 3' untranslated region
Intronic	10	Variants within an intron
Upstream	11	variant overlaps 1-kb region upstream of transcription start site (use --neargene x to change this, the unit of x is base-pair)
Downstream	12	variant overlaps 1-kb region downtream of transcription end site (use --neargene x to change this, the unit of x is base-pair)
ncRNA	13	variant overlaps a transcript without coding annotation in the gene definition (see Notes below for more explanation)
Intergenic	14	variant is in intergenic region
Unknown	15	Variants KGGSeq failed to mapp

4.4. Prioritize variants for complex disorders

Index	Arguments	Comments
Select pipeline type		
1	--prioritize-complex	Prioritize variants for complex disorders. But it has to go with the specific settings as described in the following.

		[compulsory]
Select plotting types		
2	--qqplot	Plot quantile-quantile plot for p-values in input summary results. [optional]
3	--mafplot	Plot histogram of minor allele frequencies for variants in input summary results. [optional]
Set filtration model		
4	--filter-model association	Exclude variants which are not interesting. Four values are possible: association, case-unique, control-unique and no. [compulsory]
5	--p-value-cutoff 0.01	Set p-value cutoffs for filtration. [optional]
Filtration according to gene features		
6	--db-gene refgene	See Table 4.1 for the same item
7	--gene-feature-in 0,1,2,3,4,5,…	See Table 4.1 for the same item
Filtration according to prediction		
8	--db-score dbnsfp	See Table 4.1 for the same item
9	--filter-nondisease-variant	See Table 4.1 for the same item
Prioritization according to PPI and pathway sharing		
10	--candi-list GeneSymol1, GeneSymol2,GeneSymol3	See Table 4.1 for the same item
10'	--candi-file /path/to/file	See Table 4.1 for the same item
Prioritization according to Literature		
11	--pubmed-mesh Type+1+Diabetes, Type+2+Diabetes,...	See Table 4.1 for the same item
Prioritization according to regulation information		
12	--tfbs-anno	See Table 4.1 for the same item
13	--filter-nontfbs-variant	See Table 4.1 for the same item
14	--enhancer-anno	See Table 4.1 for the same item
15	--filter-nonEnhancer-variant	See Table 4.1 for the same item

4.5. Output setting and annotation

ID	Arguments	Annotations
1	--out path/to/output/file	Specify path and prefix name of outputs. It is "kggseq" by default. [compulsory]

2	--excel	Export data in an excel file; it is a text file by default. [optional]
3	--o-polyphen	Produce inputs of Polyphen. [optional]
4	--o-seattleseq	Produce inputs of SeattleSeq. [optional]
5	--o-annovar	Produce inputs of ANNOVAR. [optional]
6	--o-vcf	Extract the VCF data of prioritized variants. [optional]
7	--o-plink	Produce inputs of PLINK, a PEDigree and a MAP file. [optional]

Each items output in the excel file are list in the following two tables, which are designed for rare mendelian disease and common complex disease respectively.

Table 4.5.1 Annotation for list of variables output for KGGSeq pipeline 1

ID	Column Name	Annotation
1	Chromsome	Chromosome number (from 1 to 22, and X, Y)
2	Position	Physical location of the variants
3	ReferenceAlternativeAllele	Reference allele and alternative allele(s)
4	GeneSymbol	Official gene symbol
5	GeneFeatures	All involved gene features of a variant Explanation of some **example** features *rs17338579:RPA1:NM_002945:intronic2* Means: This variant is in the 2nd introns of a transcript. *.:TMEM107:NM_183065:3UTR+768* Means: This variant is in the 3' UTR and is 768bp downstram from the first 3'UTR site on the 3' side. *.:CTC1:NM_025099:5UTR-42* Means: This variant is within the 5' UTR and is at 42bp upstream from the first 5'UTR site on the 5' side. *AK5:NM_174858:5splicing14-2* Means: This variant is at 2bp upstream of the 5' side of 14th exon, which is also the splicing site. *.:SASS6:NM_194292:upstream-24* Means: This variant is at 24bp upstream of this transcript.
6	GeneFeature	The most promising gene feature of a variant
7	dbAltAF	Alternative allele frequency in database
8	PhyloP_score	Predicted deleteriousness score from PhyloP
9	SIFT_score	Predicted deleteriousness score from SIFT
10	Polyphen2_score	Predicted deleteriousness score from Polyphen2
11	LRT_score	Predicted deleteriousness score from LRT (likelihood ratio test)
12	MutationTaster_score	Predicted deleteriousness score from mutationTaster
13	AffectedRefHomGtyNum	Numbers of reference homozygotes in affected sample

<table>
<tr><td>14</td><td>AffectedHetGtyNum</td><td>Numbers of heterozygotes in affected sample</td></tr>
<tr><td>15</td><td>AffectedAltHomGtyNum</td><td>Numbers of alternative homozygotes in affected sample</td></tr>
<tr><td>16</td><td>UniProtFeature</td><td>Annotate a variant of coding gene using the UniProt protein features.
<table>
<tr><th>UniProtFeature</th><th>Definition</th></tr>
<tr><td>region of interest</td><td>Extent of a region of interest in the sequence</td></tr>
<tr><td>active site</td><td>Amino acid(s) involved in the activity of an enzyme.</td></tr>
<tr><td>calcium-binding region</td><td>Extent of a calcium-binding region.</td></tr>
<tr><td>glycosylation site</td><td>Glycosylation site.</td></tr>
<tr><td>chain</td><td>Extent of a polypeptide chain in the mature protein.</td></tr>
<tr><td>coiled-coil region</td><td>Extent of a coiled-coil region</td></tr>
<tr><td>compositionally biased region</td><td>Extent of a compositionally biased region</td></tr>
<tr><td>sequence conflict</td><td>Different papers report differing sequences.</td></tr>
<tr><td>cross-link</td><td>Posttranslationally formed amino acid bonds.</td></tr>
<tr><td>disulfide bond</td><td>Disulfide bond.</td></tr>
<tr><td>DNA-binding region</td><td>Extent of a DNA-binding region.</td></tr>
<tr><td>domain</td><td>Extent of a domain, which is defined as a specific</td></tr>
<tr><td>helix</td><td>DSSP secondary structure</td></tr>
<tr><td>intramembrane region</td><td></td></tr>
<tr><td>initiator methionine</td><td>Initiator methionine.</td></tr>
<tr><td>modified residue</td><td></td></tr>
<tr><td>lipid moiety-binding region</td><td></td></tr>
<tr><td>metal ion-binding site</td><td>Binding site for a metal ion.</td></tr>
<tr><td>short sequence motif</td><td>Short (up to 20 amino acids) sequence motif of biological interest</td></tr>
<tr><td>mutagenesis site</td><td>Site which has been experimentally altered.</td></tr>
<tr><td>non-consecutive residues</td><td>Non-consecutive residues.</td></tr>
<tr><td>non-standard amino acid</td><td>Non-standard aminoacid used for pyrrolysine</td></tr>
<tr><td>non-terminal residue</td><td></td></tr>
<tr><td>nucleotide phosphate-binding region</td><td>Extent of a nucleotide phosphate binding region.</td></tr>
<tr><td>peptide</td><td>Extent of a released active peptide.</td></tr>
<tr><td>propeptide</td><td>Extent of a propeptide.</td></tr>
<tr><td>repeat</td><td>Extent of an internal sequence repetition.</td></tr>
<tr><td>signal peptide</td><td>Extent of a signal sequence (prepeptide).</td></tr>
<tr><td>site</td><td>Any interesting single amino-acid site on the sequence, that</td></tr>
</table></td></tr>
</table>

<table>
<tr><td></td><td></td><td><table>
<tr><td>strand</td><td>DSSP secondary structure</td></tr>
<tr><td>transit peptide</td><td></td></tr>
<tr><td>topological domain</td><td>Topological domain</td></tr>
<tr><td>transmembrane region</td><td>Extent of a transmembrane region.</td></tr>
<tr><td>turn</td><td>DSSP secondary structure</td></tr>
<tr><td>unsure residue</td><td>Uncertainties in the sequence</td></tr>
<tr><td>sequence variant</td><td>Authors report that sequence variants exist.</td></tr>
<tr><td>splice variant</td><td>Description of sequence variants produced by alternative</td></tr>
<tr><td>zinc finger region</td><td>Extent of a zinc finger region.</td></tr>
<tr><td>binding site</td><td></td></tr>
</table></td></tr>
<tr><td>17</td><td>GeneDescription</td><td>Description of the gene</td></tr>
<tr><td>18</td><td>Pseudogenes</td><td>Psudogenes of the variants gene. This can give users caution of possible mapping error due to these pseudogenes</td></tr>
<tr><td></td><td>SIFT_pred</td><td>SIFT_score: SIFT score, If a score is smaller than 0.05 the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)".</td></tr>
<tr><td></td><td>Polyphen2_pred</td><td>Polyphen2 prediction based on HumVar score, "D" ("porobably damaging") if it is in [0.909,1], "P" ("possibly damaging") if it is in [0.447,0.908]; and "B" ("benign") if it is in [0,0.446]. Multiple entries separated by ";".</td></tr>
<tr><td></td><td>LRT_pred</td><td>LRT prediction, D(eleterious), N(eutral) or U(nknown)</td></tr>
<tr><td></td><td>MutationTaster_pred</td><td>MutationTaster_pred: MutationTaster prediction, "A" ("disease_causing_automatic"),"D" ("disease_causing"), "N" ("polymorphism") or "P" ("polymorphism_automatic")</td></tr>
<tr><td>21</td><td>DiseaseCausalProb_ ExoVar TrainedModel</td><td>Conditional probability of disease causal given the five deleteriousness scores under a logistic regression model trained by a dataset named ExoVar http://statgenpro.psychiatry.hku.hk/limx/kggseq/download/ExoVar.xls. Details of the calculation of the conditional probability are described in Reference Paper 2. Sorry, the paper is under review now.</td></tr>
<tr><td>22</td><td>IsRareDiseaseCausal_ ExoVar TrainedModel</td><td>Whether the variant is predicted to be rare disease causal or not according to a logistic regression model trained by the ExoVar dataset. Details of methods are described in Reference Paper 2.</td></tr>
<tr><td>23</td><td>IsWithinCandidateGene</td><td></td></tr>
<tr><td>24</td><td>CandidateGeneOrPPI</td><td>Candidate gene in which a variant is within and protein-protein interactions between a candidate gene provided and the variant's gene</td></tr>
<tr><td>25</td><td>SharedPathway</td><td>Available pathways shard by the variant's gene and at least one of candidate genes provided</td></tr>
<tr><td>26</td><td>LongestIBSRegion</td><td>Longest region of IBS sharing by patients.</td></tr>
<tr><td>27</td><td>LongestIBSRegionLengthinbp</td><td>The length of the longest IBS region</td></tr>
<tr><td>28</td><td>DispensableDeleteriousVariants</td><td>Numbers of dispensable deleterious variants for the gene in 1000 Genome Project data</td></tr>
<tr><td>29</td><td>PubMedIDIdeogram</td><td>PubMed ID of papers mentioning ideogram of the variant and the disease in question in the title or abstract together.</td></tr>
<tr><td>30</td><td>PubMedIDGene</td><td>PubMed ID of papers mentioning gene symbol of the variant and the disease in question in the title or abstract together.</td></tr>
</table>

Table 4.5.2 Annotation for list of variables output for KGGSeq pipeline 2

ID	Column Name	Annotation
1	Chromsome	Chromosome number (from 1 to 22, and X, Y)
2	Position	Physical location of the variants
3	ReferenceAlternativeAllele	Reference allele and alternative allele(s)
4	GeneSymbol	Official gene symbol
5	GeneFeatures	All involved gene features of a variant Explanation of some example features *rs17338579:RPA1:NM_002945:intronic2* Means: This variant is in the 2nd introns of a transcript. *.:TMEM107:NM_183065:3UTR+768* Means: This variant is in the 3' UTR and is 768bp downstream from the first 3'UTR site on the 3' side. *.:CTC1:NM_025099:5UTR-42* Means: This variant is within the 5' UTR and is at 42bp upstream from the first 5'UTR site on the 5' side. *AK5:NM_174858:5splicing14-2* Means: This variant is at 2bp upstream of the 5' side of 14th exon, which is also the splicing site. *.:SASS6:NM_194292:upstream-24* Means: This variant is at 24bp upstream of this transcript.
6	GeneFeature	The most promising gene feature of a variant
7	PhyloP_score	Predicted deleteriousness score from PhyloP
8	SIFT_score	Predicted deleteriousness score from SIFT
9	Polyphen2_score	Predicted deleteriousness score from Polyphen2
10	LRT_score	Predicted deleteriousness score from LRT (likelihood ratio test)
11	MutationTaster_score	Predicted deleteriousness score from mutationTaster
12	AffectedHetGtyNum	Numbers of reference homozygotes in cases
13	AffectedHetGtyNum	Numbers of heterozygotes in cases
14	AffectedAltHomGtyNum	Numbers of alternative homozygotes in cases
15	UnaffectedRefHomGtyNum	Numbers of reference homozygotes in controls
16	UnaffectedHetGtyNum	Numbers of heterozygotes in controls
17	UnaffectedAltHomGtyNum	Numbers of alternative homozygotes in controls

<table>
<tr><td>18</td><td>UniProtFeature</td><td>Annotate a variant of coding gene using the UniProt protein features.
<table>
<tr><th>UniProtFeature</th><th>Definition</th></tr>
<tr><td>region of interest</td><td>Extent of a region of interest in the sequence</td></tr>
<tr><td>active site</td><td>Amino acid(s) involved in the activity of an enzyme.</td></tr>
<tr><td>calcium-binding region</td><td>Extent of a calcium-binding region.</td></tr>
<tr><td>glycosylation site</td><td>Glycosylation site.</td></tr>
<tr><td>chain</td><td>Extent of a polypeptide chain in the mature protein.</td></tr>
<tr><td>coiled-coil region</td><td>Extent of a coiled-coil region</td></tr>
<tr><td>compositionally biased region</td><td>Extent of a compositionally biased region</td></tr>
<tr><td>sequence conflict</td><td>Different papers report differing sequences.</td></tr>
<tr><td>cross-link</td><td>Posttranslationally formed amino acid bonds.</td></tr>
<tr><td>disulfide bond</td><td>Disulfide bond.</td></tr>
<tr><td>DNA-binding region</td><td>Extent of a DNA-binding region.</td></tr>
<tr><td>domain</td><td>Extent of a domain, which is defined as a specific</td></tr>
<tr><td>helix</td><td>DSSP secondary structure</td></tr>
<tr><td>intramembrane region</td><td></td></tr>
<tr><td>initiator methionine</td><td>Initiator methionine.</td></tr>
<tr><td>modified residue</td><td></td></tr>
<tr><td>lipid moiety-binding region</td><td></td></tr>
<tr><td>metal ion-binding site</td><td>Binding site for a metal ion.</td></tr>
<tr><td>short sequence motif</td><td>Short (up to 20 amino acids) sequence motif of biological interest</td></tr>
<tr><td>mutagenesis site</td><td>Site which has been experimentally altered.</td></tr>
<tr><td>non-consecutive residues</td><td>Non-consecutive residues.</td></tr>
<tr><td>non-standard amino acid</td><td>Non-standard aminoacid used for pyrrolysine</td></tr>
<tr><td>non-terminal residue</td><td></td></tr>
<tr><td>nucleotide phosphate-binding region</td><td>Extent of a nucleotide phosphate binding region.</td></tr>
<tr><td>peptide</td><td>Extent of a released active peptide.</td></tr>
<tr><td>propeptide</td><td>Extent of a propeptide.</td></tr>
<tr><td>repeat</td><td>Extent of an internal sequence repetition.</td></tr>
<tr><td>signal peptide</td><td>Extent of a signal sequence (prepeptide).</td></tr>
<tr><td>site</td><td>Any interesting single amino-acid site on the sequence, that</td></tr>
<tr><td>strand</td><td>DSSP secondary structure</td></tr>
<tr><td>transit peptide</td><td></td></tr>
<tr><td>topological domain</td><td>Topological domain</td></tr>
<tr><td>transmembrane region</td><td>Extent of a transmembrane region.</td></tr>
<tr><td>turn</td><td>DSSP secondary structure</td></tr>
<tr><td>unsure residue</td><td>Uncertainties in the sequence</td></tr>
<tr><td>sequence variant</td><td>Authors report that sequence variants exist.</td></tr>
<tr><td>splice variant</td><td>Description of sequence variants produced by alternative</td></tr>
<tr><td>zinc finger region</td><td>Extent of a zinc finger region.</td></tr>
<tr><td>binding site</td><td></td></tr>
</table></td></tr>
<tr><td>19</td><td>P</td><td>Individual variant p-value in allelic model</td></tr>
</table>

20	PDOM	Individual variant p-value in dominant model
21	PREC	Individual variant p-value in recessive model
22	GeneDescription	Description of the gene
23	Pseudogenes	Psudogenes from psudogene database
	SIFT_pred	SIFT_score: SIFT score, If a score is smaller than 0.05 the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)".
	Polyphen2_pred	Polyphen2 prediction based on HumDIV score, "D" ("porobably damaging") if it is in [0.957,1], "P" ("possibly damaging") if it is in [0.453,0.956]; and "B" ("benign") if it is in [0,0.452]. Multiple entries separated by ";".
	LRT_pred	LRT prediction, D(eleterious), N(eutral) or U(nknown)
	MutationTaster_pred	MutationTaster_pred: MutationTaster prediction, "A" ("disease_causing_automatic"),"D" ("disease_causing"), "N" ("polymorphism") or "P" ("polymorphism_automatic")
24	DiseaseCausalProbHumDivTrainedModel.	Conditional probability of disease causal given the five deleteriousness scores. Details of the calculation of the conditional probability will be described in the Reference Paper 2. Sorry, the paper is under review now.
25	IsComplexDiseaseCausalHumDivTrainedModel	Whether the variant is predicted to be complex disease causal or not according to the conditional probability.
26	IsWithinCandidateGene	Candidate gene within which a variant is
27	PPI	Protein-protein interactions between a candidate gene provided and the variant's gene
28	SharedPathway	Available pathways shard by the variant's gene and at least one of candidate genes provided
29	TFBSconsSite [tfbsName:rawScore:zScore]	Available tfbs regions where the variants is located.
30	vistaEnhancer [enhancerName: positive/negative]	Available enhancer regions where the variants is located.
31	PubMedIDIdeogram	PubMed ID of papers mentioning ideogram of the variant and the disease in question in the title or abstract together.
32	PubMedIDGene	PubMed ID of papers mentioning gene symbol of the variant and the disease in question in the title or abstract together.

5. FAQs

1) **Does Kggseq provide gene expression information in the output?**
   Kggseq does not implement any gene expression database for annotating location or quantity of gene expression. But users can input Kggseq's final gene list to *GeneCards* for gene expression information.

2) **Why Kggseq does not read my VCF file?**
   If you use standard VCF output from GATK pipeline, it usually contains variants

on mitochondrial DNA. However, mitochondrial DNA is not annotated by gene feature database. Therefore Kggseq currently only accept VCF file excluding variants on ChrM.

3) **Is there a link for annotating shared pathways in terms of biological function?**
Kggseq implements MsigDB database, therefore each pathway item output by Kggseq can be directed to a weblink which contain concrete information for the corresponding pathway.

4) **How shall I prepare the candidate gene list? What is the gene list for?**
The candidate gene list can include any gene interesting to the user. It may be verified causal gene for the disease, or potential causal gene. The gene list is used to fish more hidden interesting genes in order to narrow down the searching region. If no gene list provided, Kggseq can still perform routine 'filtration + sharing' strategy for sequencing data.

5) **Suppose I have in-house control genomes in VCF format, how can I transform it into Kggseq acceptable file?**
There are two alternative ways to allow you use the in-house VCF data (containing genotypes of multiple individuals) as filter.
   1. Convert the VCF data into a standard kggseq local filter format by the command: *--make-filter --vcf-file path/to/vcffile --buildver-in hg18 --out test.hg19.var --buildver-out hg19* first. And then set the --local-filter *test.hg19.var*
   2. Directly set *--local-filter-vcf path/to/vcffile.* However, this way is less efficient than the above way as it will take additional time to parse the VCF data.

6) **What is the default setting for minor allele frequency (MAF) cut-offs in the filtration step? Is there a practical guidance for MAF selection?**
The default setting for MAF cut-offs is 0, which means only novel variants can survive in the filtration process. MAF selection is a difficult problem, but generally two suggestions can be adopted: for rare Mendelian disease, MAF cut-off at 0 or 0.01 is reasonable; for common complex disease of relatively rare alleles with large effect size, MAF cut-off is dependent on disease prevalence, penetrance and genetic model, so a higher cut-off like 0.05 might be more reasonable.

7) **Can I use ANNOVAR and Kggseq together on my dataset?**
Kggseq is quite flexible for interacting with other sequence-oriented analytical programs/software (including SOAPSNP, ANNOVAR, PLINK/seq, VAAST, etc). It can accept various input formats, and output different kinds of sequence data. In case of ANNOVAR, Kggseq can read ANNOVAR-formatted sequence variants, and write the final remaining variants in ANNOVAR format.

8) **Can I run Kggseq on my Lenovo or Mac laptop? Is it time consuming to run a complete Kggseq process?**
Normally, Kggseq run well and fast with >=1 GB RAM memory. Hence current laptop are certainty affordable for running Kggseq. The whole process need only

<10 minutes, unless PubMed searching is set in the parameter file. For PubMed searching, it really depends on the speed of the network and also the loading of the NCBI PubMed database.

9) How do I report an error or bugs to Kggseq?

You are welcomed to join our google group (http://groups.google.com/group/kggseq_user?hl=en). This site is used for communication and discussion of Kggseq usage and functions. You can also directly write an email to limx54@yahoo.com, or kggseq_user@googlegroups.com.