**Next generation sequencing (NGS) data analysis for genetic studies of rare diseases
—The use of KGGSeq to filter, prioritize and annotate sequence variants for identifying disease-causal mutations**

*By Miaoxin Li*

**The best way to learn how to use KGGSeq is to go through our online manual** <http://statgenpro.psychiatry.hku.hk/limx/kggseq/doc/UserManual.html> !!!

**Tasks:**

|  |
| --- |
| **Task 1:** ***Predict pathogenic non-synonymous variants in CASP12 gene***1. Copy VCF files in *miaoxin/2015/KGGSeq/examples* into your folder.
2. Open your terminal and enter the folder, *examples/genes*.
3. Use KGGSeq to annotate the sequence variants and predict the variants run “*kggseq --no-gty-vcf-file ./CASP12.vcf --db-gene refgene,gencode,knowngene --db-score dbnsfp --mendel-causing-predict all --excel*”
4. View the excel file, *kggseq.flt.xlsx*. And close it.
5. Filter out common variants with MAF (>0.01) “*kggseq --no-gty-vcf-file ./CASP12.vcf --db-gene refgene,gencode,knowngene --db-score dbnsfp --mendel-causing-predict all --excel --db-filter 1kg201204,* *exac --rare-allele-freq 0.01*”

*Run these commands when it is on your computer.*java *-*Xmx1g *-*jar *kggseq.jar --no-lib-check --no-resource-check --no-gty-vcf-file genes/CASP12.vcf --db-gene refgene,gencode,knowngene --db-score dbnsfp --mendel-causing-predict all --excel*java *-*Xmx1g *-*jar *kggseq.jar --no-lib-check --no-resource-check --no-gty-vcf-file genes/CASP12.vcf --db-gene refgene,gencode,knowngene --db-score dbnsfp --mendel-causing-predict all --excel --db-filter 1kg201204,* *exac --rare-allele-freq 0.01* |
| **Questions:** 1. How many non-synonymous variants are this gene has in the VCF file?2. How many non-synonymous variants are predicted to be pathogenic?3. How many **rare** non-synonymous variants are predicted to be pathogenic? |

|  |
| --- |
| **Task 2:** ***Identify sequence variant candidate that may cause dominant Arthrogryposis***1. Open your terminal and enter the folder, *examples*.
2. Take a look at the input vcf and pedigree file by typing “*less rare.disease.hg19.vcf*” and “*less rare.disease.ped.txt*”
3. Identify the causal mutation by typing “*kggseq ---vcf-file rare.disease.hg19.vcf --ped-file rare.disease.ped.txt --out arh --excel --genotype-filter 3,4,5,6 --seq-qual 50 --seq-mq 20 --seq-fs 60 --gty-qual 20 --gty-dp 8 --db-gene refgene --gene-feature-in 0,1,2,3,4,5 --db-filter-hard dbsnp138nf --db-filter exac,1kg201204,dbsnp138,dbsnp141,ESP6500AA,ESP6500EA --rare-allele-freq 0.01 --db-score dbnsfp --mendel-causing-predict all --filter-nondisease-variant --superdup-filter --gene-var-filter 4 --genome-annot --omim-annot --candi-list ECEL1,MYBPC1,TNNI2,TNNT3,TPM2 --ppi-annot string --ppi-depth 1 --pathway-annot cura --pubmed-mining Arthrogryposis,Arthrogryposis+multiplex+congenita*”
4. View the excel file, *arh.flt.xlsx*

**Note**: Go to the short tutorial page (http://statgenpro.psychiatry.hku.hk/limx/kggseq/doc/MendelianDisease.htm) to learn the combination of multiple options.*Run these commands when it is on your computer.*java *-*Xmx1g *-*jar *kggseq.jar --no-lib-check --no-resource-check --vcf-file examples/rare.disease.hg19.vcf --ped-file examples/rare.disease.ped.txt --out arh --excel --genotype-filter 3,4,5,6 --seq-qual 50 --seq-mq 20 --seq-fs 60 --gty-qual 20 --gty-dp 8 --db-gene refgene --gene-feature-in 0,1,2,3,4,5 --db-filter-hard dbsnp138nf --db-filter exac,1kg201204,dbsnp138,dbsnp141,ESP6500AA,ESP6500EA --rare-allele-freq 0.01 --db-score dbnsfp --mendel-causing-predict all --filter-nondisease-variant --superdup-filter --gene-var-filter 4 --genome-annot --omim-annot --candi-list ECEL1,MYBPC1,TNNI2,TNNT3,TPM2 --ppi-annot string --ppi-depth 1 --pathway-annot cura --pubmed-mining Arthrogryposis,Arthrogryposis+multiplex+congenita* |
| **Questions:** 1. How many variants were filtered out according to quality control?2. How many variants were filtered out according to genetic model?3. How many common variants were filtered out?4. What are promising evidences to support the causality of these variants? |