LAB SESSION 3 -- VARIANT ANNOTATION AND FILTERING

I. Running Full CLCG Pipeline in Galaxy (TO BE DONE TOGETHER)

1. Create a new history (Options → Create New) and give it a name
2. Get FASTQ sequencing read files:
    a. Shared Data → Data Libraries → IIHG Bioinformatics Course
    b. Sample 1 → Check boxes next to FASTQ files (1_R1.fastq, 1_R2.fastq)
    c. Click "Go" to import data sets to current history
3. Get OtoSCOPE and dbSNP files:
    a. Shared Data → Data Libraries → IIHG Bioinformatics Course → check box next to OtoSCOPEv4.bed → Click "Go" to import data sets to current history
    b. Shared Data → Data Libraries → CLCG Pipeline Design and Reference Files → check box next to dbSNP_132.hg19.sort2.vcf → Click "Go" to import data sets to current history
4. Import CLCG Pipeline:
    a. Shared Data → Published Workflows → click CLCG Illumina Paired-End Workflow → click the green plus "Import Workflow" button
5. Click "Analyze Data" to return to home window:
    a. You should have four data sets in your history: 1_R1.fastq, 1_R2.fastq, OtoSCOPEv4.bed, dbSNP reference
6. Run workflow:
    a. On left side of screen click "workflows" → "All Workflows"
    b. Click "Imported: CLCG Illumina Paired End Workflow"
    c. Name your sample
    d. Select forward FASTQ file: pick "1_R1.fastq" from drop-down menu
    e. Select reverse FASTQ file: pick "1_R2.fastq" from drop-down menu
    f. Select design file (Bed Format): pick "otoscopev4.bed" from drop-down menu
    g. Select dbSNP_REF file: pick "dbSNP reference" from drop-down menu
    h. Scroll to very bottom and click "run workflow"
    i. That's it! Just wait for it to finish.

II. Pathogenic Variant Identification in Sample 1

See Part IV below if you do not have the annotated Excel file available -- just import the pre-annotated file before continuing

1. Download Annotated Variant File:
    a. When workflow is done, click the Save button (floppy disk) next to sample name.annotated..xls file to download
    b. Find the file on your computer and open in Excel
2. Add a new quality score in Excel -- Quality divided by depth (Q/D):
    a. click in the "Q_GT" column
    b. Insert → column
    c. Name the column (1st row) "QD"
    d. Click cell H2 and type "=", click cell G2, type "/", click cell F2, then ENTER (your QD should be calculated)
    e. Copy and paste this cell to all rows in this column
3. Filtering and Prioritizing in Excel
    a. Click once in any cell containing data
    b. Data → filter (you should see arrows appear on top row)
    c. Click arrow to filter by column
    d. Filter for depth (#COVERED > or = to 10)
    e. Filter for quality (QVAR > or = to 30)
    f. Filter for quality (QD > 10)
    g. Filter out synonymous and intronic changes (SYNONYMOUS column -- uncheck "synonymous" and "blank")
    h. Filter our high MAF variants (Minor Allele Frequency column -- uncheck all but "0", "0.02225", "0.2213" and "blank") ***make sure to include blanks (these are variants without ANY MAF)
4. Questions:
    a. How many rows do you have?
    b. How many UNIQUE variants (i.e. variants of unknown significance) do you have?
    c. What are your candidate variants for an Autosomal Recessive inheritance?
    d. What is your diagnosis? Check http://deafnessvariationdatabase.org to confirm your result.

III. Annotate with wANNOVAR and compare to CLCG annotation

1. Download your VCF file:

    a. Go back to your history in Galaxy
    b. Click the Save button (floppy disk) next to samplename.unifiedGenotype.vcf file to download
2. Submit file to wANNOVAR:
    a. Go to http://wannovar.usc.edu
    b. Put in a sample identifier and your email address
    c. Input file name: click “choose file” and browse to find your VCF file
    d. Leave all other fields default (VCF file, hg19, refseq genes, no filtering) and click “submit” -- wait for it to finish
3. Compare Annotations:
    a. Download “Genome Summary” file from ANNOVAR
    b. Open in excel
    c. Compare the variants at these positions between CLCG and ANNOVAR annotation:
        i. chr5:90459600
        ii. chr1:109472773

IV. Samples 2, 3, and 4

1. Instead of running the entire pipeline (or in case you can’t get it to work) several files are available in Shared Data → Data Libraries → IIHG Bioinformatics Course → Lab Session 3 → Annotated OtoSCOPE files:
    a. VCF file (can be annotated in galaxy -- see below)
    b. ANNOVAR annotated file (.annovar.csv): can be opened in excel for filtering
    c. SeattleSeq annotated file (.seattleseq.csv): can be opened in excel for filtering
2. Using the CLCG annotation tool in Galaxy to annotate a VCF file:
    a. Import the VCF file to your history (Shared Data → Data Libraries → IIHG Bioinformatics Course → Lab Session 3 → Annotated OtoSCOPE files)
    b. Under tools click “CLCG Annotation Pipeline”
    c. Download excel file when done
3. Use your method of choice to identify the causative variants (if any)!

V. Exome data

1. Create a new history
2. Import the exome data sets in Shared Data → Data Libraries → IIHG Bioinformatics Course → Lab Session 3
3. Run the CLCG Annotation tool as in IV.2 above OR submit to wANNOVAR for analysis OR SeattleSeq
4. If things aren’t working, just use the already annotated data sets in hared Data → Data Libraries → IIHG Bioinformatics Course → Lab Session 3 → Annotated Exome Data
5. Family 1: Filter to just look in the linked region (chr5:104,881,398-180,503,151) for autosomal recessive causative variant
6. Family 2: Filter for a unique (never seen before), heterozygous non-synonymous variant that is VERY conserved and damaging and shared between both affected individuals