getting_started_with_ISB-google-170217.pdf
Shared resource for the PanCanAtlasGermline group

Essential links

Login page for Accessing TCGA Controlled Data through link to authentication page: ISB-CGC3

Dashboard:
https://console.cloud.google.com/home/dashboard?project=isb-cgc-06-0004&pli=1

Jay's getting started on google cloud doc:

Sheila's more advanced getting started on google cloud doc:
https://docs.google.com/document/d/1f1YBVG1dAhpF-Un5lp70kMI8Yo2rD3NNQdpl0FDAu-c/edit?usp=sharing

"You might also find some of the tutorial material on the ISB-CGC helpful in beginning to work in the Google Cloud. In particular there are quick tutorials on Google BigQuery, the Google Cloud Console, and Cloud Shell (none of that is ISB-CGC centric in anyway and just introduces you to the Google Cloud). Complete Google documentation is, of course, the final and best reference."

PCA Germline Working Group Release v.1.0 (workflow to generate such files available on github: downstream steps are mostly in https://github.com/ding-lab/PanCanAtlasGermline/blob/master/analysis/process_files/germline/local/work.log.sh):

gs://dinglab/isb-cgc/tcga/germline/release 1.1/:

all/: pass-filtered VCFs of 10389 samples merged into a unified compressed VCF; also contained tumor barcodes for TN-swap samples

PCA.r1.TCGAbarcode.merge.tnSwapCorrected.10389.vcf.gz
PCA.r1.TCGAbarcode.merge.tnSwapCorrected.10389.vcf.gz.tbi
Tn_swap_sample_list.txt

combineClassified/:

a. **PCA_pathVar_integrated_filtered_adjusted.tsv**
Merged with multiple attributes and classified based on ClinVar, database, CharGer classification and some post-hoc prioritization (use Overall_Classification column).
This is the final set of rare, prioritized cancer variants now used the exon manuscript.

gs://dinglab/isb-cgc/tcga/germline/release 1.0/

sample_list/pca_table.20171019.wclin.tsv: a table containing the UUID, TCGA barcode and original file paths of the 10,467 prioritized, white-list, clinical-data available sample (best bam file for that patient) with final set of VCFs: originated from **[pca_table.20171019.tsv.zip]**: 10,568 samples with available variant calls

individualVCF/: each white-list sample has one filtered compressed VCF ({sampleUUID}.AD.5.vcf.gz).
TOTAL: 10568 objects, 69088267902 bytes (64.34 GiB).
readme pending to describe QC for these samples; and the final VCFs merged.

batchMerge/: each batch (300 VCFs) of individual VCFs merged into compressed VCFs PCA_aa to PCA_bj.

all/: pass-filtered VCFs of 10568 samples merged into a unified compressed VCF

1. **unannotated/**: one unannotated indexed VCF (28GB).
   a. Not sorted because exceeding memory limits on the largest VMs on Gcloud. Help sorting would be largely appreciated.

2. **annotated/:**
    a. **PCA.germline.chr{i}.snpeff.snpsift.vcf.gz**
    Snpeff and snpsift annotated vcfs - contains union of calls no genotype information
    b. **PCA.germline.chr{i}.annotated_variants.csv.gz**
    GWAVA annotations intermediate file with annotations used in GWAVA Random Forest model **non-coding only**
    c. **PCA.germline.chr{i}.GWAVA_scores.tsv**
    GWAVA scores for non-coding variants

exonOnly/: pass-filtered VCFs of 10568 samples merged into a unified compressed VCF, exon region only. Sorted.

1. **unannotated/**: one unannotated indexed VCF (4.7GB) containing all samples' genotype.
2. **annotated/:** VEP annotated variants of each chromosome's VCFs without the sample genotype.
3. **CharGer_classified/**: CharGer classified likely pathogenic/pathogenic variants (49,123 variants).
4. **validated/**:
    a. charged.PCA.r1.TCGAbarcode.merge.exon.ALL.vcf.samples.cleaned.expanded.AFcorrected.lowAF.sele.labeled.tsv.gz
    All rare pathogenic variants validated by readcounts (31,963 variants).
    b. charged.PCA.r1.TCGAbarcode.merge.exon.ALL.vcf.samples.cleaned.expanded.AFcorrected.lowAF.sele.labeled.rare.cancer.tsv
    Rare, pathogenic cancer-variants validated by readcounts (1,982 variants).
5. **reviewed/**:
    a. charged.PCA.r1.TCGAbarcode.merge.exon.ALL.vcf.samples.cleaned.expanded.AFcorrected.lowAF.sele.labeled.rare.cancer.pass.tsv
    1,476 final rare, cancer-related pathogenic variants.
    b. Germline_PCA_RJ_manualReview.zip
    Reyka's folder of manual review results including IGV shots for individual variants.
6. **combineClassified/**:
    a. **PCA_pathVar_integrated_filtered.tsv**
    Merged with multiple attributes and classified based on ClinVar, database and CharGer classification (use Overall_Classification column).
    This is NO LONGER the final set of rare, prioritized cancer variants used the exon manuscript.
    b. **PanCan_ClinicalData_V4_wAIM_filtered10389.txt**
    In the end we only retained 10,389 samples that passed QC (showing good coverage, no outlying numbers of variants called, and high concordance with SNP array data). The clinical table of the associated 10,389 cases is here based on Clinical data from the clinical working group and ethnicity calls by the AIM working group.

Resources on Google

ExAC Database r0.3.1

gs://gnomad-public/legacy/exacv1_downloads/release0.3.1/ExAC.r0.3.1.sites.vep.vcf.gz
gs://gnomad-public/legacy/exacv1_downloads/release0.3.1/ExAC.r0.3.1.sites.vep.vcf.gz.tbi
gs://gnomad-public/legacy/exacv1_downloads/release0.3.1/subsets/ExAC_nonTCGA.r0.3.1.sites.vep.vcf.gz
gs://gnomad-public/legacy/exacv1_downloads/release0.3.1/subsets/ExAC_nonTCGA.r0.3.1.sites.vep.vcf.gz.tbi

ExAC Database r1

gs://gnomad-public/exac/ExAC.r1.sites.vep.vcf.gz
gs://gnomad-public/exac/ExAC.r1.sites.vep.vcf.gz.tbi

gs://dinglab/reference/ExAC.r1/subsets/ExAC_nonTCGA.r1.sites.vep.vcf.gz
gs://dinglab/reference/ExAC.r1/subsets/ExAC_nonTCGA.r1.sites.vep.vcf.gz.tbi

Data generated by the project

Big query

1. https://bigquery.cloud.google.com/table/isb-cgc-06-0004:ExAC.ExAC_nonTCGA_r0_3_1
    a. ExAC nonTCGA VCF
2. https://bigquery.cloud.google.com/table/isb-cgc-06-0004:tes ting.germline_vcf_14403
    a. Concatenated VCF of combined germline calls for 14,403 samples with UUIDs
    #DO NOT USE (may be erroneous in merged format/info fields)

Storage

Note: many TCGA data source syncs with other working group. For example, the miRNA data and batch-corrected RNA-Seq expression data are downloaded from PanCanAtlas groups (https://www.synapse.org/#!Synapse:syn4557014). They also have other data like methylation data. We also downloaded the clinical data from the clinical data working group.

gs://dinglab/isb-cgc/tcga/

1. **gs://dinglab/isb-cgc/tcga/reference_files/**
   Files used for processing/reference
    a. **gs://dinglab/isb-cgc/tcga/reference_files/20160713_Rahman_KJ_KH_gene_table.txt**: a curated table of 152 cancer predisposition genes extending from the 114 Rahman 2014 gene list by Kim Johnson and Kuan-lin Huang, including their respective cancer types and mode of inheritance.
    b. **gs://dinglab/isb-cgc/tcga/reference_files/20160713_Rahman_KJ_KH_gene_table_list.txt**: the above list with only gene names
    c. **gs://dinglab/isb-cgc/tcga/reference_files/ExAC.r1.sites.vep.vcf.gz**: ExAC.r1 VCF
    d. **gs://dinglab/isb-cgc/tcga/reference_files/ExAC.r1.sites.vep.vcf.gz.tbi**: index file for ExAC.r1 VCF
    e. **gs://dinglab/isb-cgc/tcga/reference_files/ROI_MultiCell_perid.txt**: ENCODE defined regulatory regions based on epigenomic signatures
    f. **gs://dinglab/isb-cgc/tcga/reference_files/Volgestin2013Science_125genes.txt**: 125 cancer genes described in Volgestin 2013 review. One should use the 152 gene list for the purpose of the germline project but this list provide classification of tumor suppressors and oncogenes that can be useful.
    g. **gs://dinglab/isb-cgc/tcga/reference_files/all_CDS_and_ncRNA_24Chroms_Contigs_1BasedStart_2bpFlanks_ForMusic**: bed files containing the exon region defined by full length transcripts obtained from in Ensembl release 70. In addition to the coding regions, the two base pairs flanking each exon that cover splice donor/acceptor sites were included and annotated with reference to the distance to the nearest coding exon.
    h. **gs://dinglab/isb-cgc/tcga/reference_files/all_CDS_ncRNA_ENCODE_multicell_ROI.bed**: ENCODE defined regulatory regions based on epigenomic signatures converted to BED file format.
        i. **gs://dinglab/isb-cgc/tcga/reference_files/all_CDS_ncRNA_ENCODE_multicell_ROI.ENSRs.bed**: Lists only the regulatory features (ENSR#)
        ii. **gs://dinglab/isb-cgc/tcga/reference_files/all_CDS_ncRNA_ENCODE_multicell_ROI.1based.rmdup.bed:** version of parent list of regions with duplicates removed
    i. **gs://dinglab/isb-cgc/tcga/reference_files/mirBase-v21-grch37.bed**: List of 1,870 microRNA precursor (hairpin) sequences retrieved from miRBase v21 database. The genomic coordinates of 1,881 miRNAs from miRBase in GRCh38 assembly were converted to GRCh37 coordinates using UCSC LiftOver tool.

2. **gs://dinglab/isb-cgc/tcga/analysis_files/**

[Files generated by analysis]

i. **gs://dinglab/isb-cgc/tcga/analysis_files/charged.ExAC.r1.sites.vep.biallelic.combine.exon.all.patho.expanded.tsv**
Pathogenic exon variants called by CharGer in ExAC.

ii. **gs://dinglab/isb-cgc/tcga/analysis_files/plink_dist_rel.tar**
IBD distrance and relatedness calculated by plink for all samples based on the genotype files

iii. **gs://dinglab/isb-cgc/tcga/analysis_files/plink_pca.tar**
Principal components (eigen values and vectors) calculated by plink for all samples based on the genotype files

iv. **gs://dinglab/isb-cgc/tcga/analysis_files/Expression_outlier_scores.tar.gz**
Regular and composite Z-scores of BoxCox transformed expression data. (Composite means that, these zscores are calculated after several normalizations by housekeeping genes)

v. **gs://dinglab/isb-cgc/tcga/analysis_files/ExAC.r1.sites.vep.biallelic.combine.fisher.anno.v2.tsv.gz**
Single variant association test results (Fisher's exact test) for all ExAC variants using ExAC nonTCGA adjusted AC/AN vs. TCGA adjusted AC/AN (inferred, see github code).

vi. **gs://dinglab/isb-cgc/tcga/analysis_files/ExAC.r1.sites.vep.biallelic.combine.fisher.anno.v2.152gene.tsv**
Single variant association test results (Fisher's exact test) for all ExAC variants using ExAC nonTCGA adjusted AC/AN vs. TCGA adjusted AC/AN (inferred, see github code). Limited to 152 predisposition genes.

vii. **gs://dinglab/isb-cgc/tcga/analysis_files/ExAC.r1.sites.vep.biallelic.combine.fisher.[ethnicity].tsv.gz**
Single variant association test results (Fisher's exact test) for all ExAC variants using ExAC nonTCGA adjusted AC/AN vs. TCGA adjusted AC/AN (inferred, see github code) in each ExAC-defined ethnicity.

viii. **gs://dinglab/isb-cgc/tcga/analysis_files/ExAC.r1.sites.vep.biallelic.combine.fisher.[ethnicity].152gene.tsv**
Single variant association test results (Fisher's exact test) for all ExAC variants using ExAC nonTCGA adjusted AC/AN vs. TCGA adjusted AC/AN (inferred, see github code) in each ExAC-defined ethnicity. Limited to 152 predisposition genes.

1. **gs://dinglab/isb-cgc/tcga/analysis_files/CharGer_input/**
    a. **20160301_Rahman_KJ_KH_gene_table_CharGer.txt**
    Curated list of 152 cancer predisposition genes with their affected tumor types and mode of inheritance.
    b. **MC3.noHypers.mericUnspecified.d10.r20.v114.clusters**
    Somatic mutations hotspot identified in MC3 cancer samples using HotSpot3D.
    c. **Clinvar_alleles.single.b37.tsv.gz**
    ClinVar alleles compiled from MacArthur lab code base.
    d. **emptyRemoved_20160428_pathogenic_variants_HGVSg_VEP.vcf**
    VCF of curated cancer predisposition variants from various databases

ix. **gs://dinglab/isb-cgc/tcga/analysis_files/WES_SNP6_array_exon_concordance.tsv**
Concordance analysis of WES calls and SNP6 array calls for only samples with high SNP6 array quality. Concordance analysis based on exon calls only.

ix. **gs://dinglab/isb-cgc/tcga/analysis_files/segregation_analysis/**
    a. **relativesFile.readme.txt**
    Readme file explaining how the relatives file below is generated
    b. **TCGA_z1_z2_relatives.tsv**
    Family pairs showing segregation

c. **TCGA_z1_z2_relatives.tsv.PCA.r1.TCGAbarcode.merge.segragatingVar.10389.tsv.gz**
Variants showing segregation with less than 1% AF

d. **TCGA_z1_z2_relatives.tsv.PCA.r1.TCGAbarcode.merge.segragatingVar.10389.AC10.tsv.gz**
Variants showing segregation with less than 0.1% AF

3. **gs://dinglab/isb-cgc/tcga/clinical**
   a. **gs://dinglab/isb-cgc/tcga/clinical/PanCan_ClinicalData_V4_20170428.txt**
   Clinical data used by the clinical group. We default to this one unless you have specific reasons to use the other one.
   b. **gs://dinglab/isb-cgc/tcga/clinical/clinical_PANCAN_patient_with_followup.tsv**
   Clinical data used by the MC3/drivers group.

4. **gs://dinglab/isb-cgc/tcga/CNV/**

5. **gs://dinglab/isb-cgc/tcga/somatic**
   a. **mc3.v0.2.8.PUBLIC.maf.gene_vclass_HGVSp_sample.gz**
   Column 1,9,16,37 of the MC3 maf containing Hugo_Symbol, Variant_Classification, Tumor_Sample_Barcode, HGVSp_Short information. The full MC3 pubic maf is here: https://www.synapse.org/#!Synapse:syn7214402/files/
   MC3 controlled access MAF (https://www.synapse.org/#!Synapse:syn5917256)

6. **gs://dinglab/isb-cgc/tcga/RNASeqExp**
   a. **EB++AdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv**
   Batch normalized RNA-Seq expression data by MC3 group
   (source: https://www.synapse.org/#!Synapse:syn4976369.3
   Batch normalization description: https://www.synapse.org/#!Synapse:syn4976363)

7. **gs://dinglab/isb-cgc/tcga/miRNA/**
   a. **bcgsc.ca_PANCAN_IlluminaHiSeq_miRNASeq.miRNAExp_whitelisted.tsv**
   miRNA expression data by MC3 group
   (https://www.synapse.org/#!Synapse:syn4557787.8)

8. **gs://dinglab/isb-cgc/tcga/RPPA/**
   a. **TCGA-RPPA-pancan-clean.txt**
   Batch-normalized RPPA data for 7754 TCGA samples (source: https://www.synapse.org/#!Synapse:syn4216793.3).

9. **gs://dinglab/isb-cgc/tcga/genotyping/**
   a. SNP genotype files
   b. **gs://dinglab/isb-cgc/tcga/genotyping/{cancer}/**
   VCF files for each sample's genotype SNPs
   c. **gs://dinglab/isb-cgc/tcga/genotyping/merge/**
   Merged VCF for each cancer
   d. **gs://dinglab/isb-cgc/tcga/genotyping/tarballs/**
   Original tarballs for each sample's VCF

10. **gs://dinglab/isb-cgc/tcga/sampleQC**
    a. Best WGS bams for tumor and normal

gs://dinglab/isb-cgc/tcga/sampleQC/WGS/TCGA_WGS_gspath_WWL_Mar2018.txt

b. Whitelisting of samples (N = 9401)
gs://dinglab/isb-cgc/tcga/sampleQC/GCS_listing.02jun2016.tcga.all.DNA.WXS.normals.tsv.noDupBam.fastqcMaxFail_2_exclKmerCont.minGoodCvg20x.analysisID_w_barcodes

c. Blacklisting components
   i. Fails coverage threshold (< 20x) (N = 6)
   gs://dinglab/isb-cgc/tcga/sampleQC/WXS_normals.all.goodCvg_20.blacklist.analysisID_w_barcodes
   ii. Fails FastQC (>= 3 tests, regardless of Kmer status) (N = 51)
   gs://dinglab/isb-cgc/tcga/sampleQC/WXS_normals.fastqc.nfail_gt_2.blacklist.analysisID_w_barcodes

11. gs://dinglab/isb-cgc/tcga/germline/

a. Germline variant calls in VCF formats
   i. gs://dinglab/isb-cgc/tcga/germline/production/{cancer}/{sample_UUID}/{caller}/
      1. gs://dinglab/isb-cgc/tcga/germline/production/{cancer}/{sample_UUID}/combine/
         a. The combined calls of all three callers (prefilter.snp_indel.vcf.*)
         b. The MAF- and ROI-filtered combined calls (prefilter.snp_indel.annotated.*)
   ii. Logs for combining the calls
      1. gs://dinglab/isb-cgc/tcga/germline/combine/
   iii. Unzipped VCF of the combined calls
      1. gs://dinglab/isb-cgc/tcga/germline/unzip/{cancer}/{sample}/combine/
   iv. Merged VCFs (based on ISB-CGC 02jun2016 manifest)
      1. PCA Germline Working Group Release v.0.9.0:
      gs://dinglab/isb-cgc/tcga/germline/release0.9/: used for non-coding variant analysis
      Exon only and filtered for N reference:
      gs://dinglab/isb-cgc/tcga/germline/release0.9/exonFiltered/: used for coding variant analysis
      Sample list (9401 sample-unique bams):
      gs://dinglab/isb-cgc/tcga/germline/release0.9/sample_lists/: a list of 9,401 pass-QC sample (uuid and TCGA barcode)
      2. Positions not in ExAC
         a. All samples called, annotated (analysisIDs applied)
            i. gs://dinglab/isb-cgc/tcga/germline/production/merge_v2/not-in-exac/3.annotated/all_list/by_analysisID/PCA.merge.analysisID.chr*.anno.vcf.gz
         b. Whitelisted samples called, annotated
            i. ***By analysisID ::*** gs://dinglab/isb-cgc/tcga/germline/production/merge_v2/not-in-exac/3.annotated/whitelisted/by_analysisID/PCA.merge.analysisID.chr*.anno.whitelist.vcf.gz
            ii. ***By TCGA barcode ::*** gs://dinglab/isb-cgc/tcga/germline/production/merge_v2/not-in-exac/3.annotated/whitelisted/by_tcgaBarcode/PCA.merge.tcgaBarcode.chr*.anno.whitelist.vcf.gz
            iii. ***TSV file of whitelisted samples ::*** gs://dinglab/isb-cgc/tcga/germline/production/merge_v2/not-in-

exac/3.annotated/whitelisted/GCS_listing.02jun2016.tcga.all.DNA.WXS.normals.tsv.noDupBam.fastqcMaxFail_2_exclKmerCont.minGoodCvg20x.analysisID_w_barcodes

3. **Positions in ExAC ("ExAConly")**
    a. All samples called, merged across cancer type
        i. Unannotated :: gs://dinglab/isb-cgc/tcga/germline/production/merge.ExAConly/all/PCA.merge.ExAConly.analysisID.vcf.gz*
        ii. Annotated ::
            1. By chromosome :: gs://dinglab/isb-cgc/tcga/germline/production/merge.ExAConly/all/annotated.by_chr/PCA.merge.ExAConly.analysisID.chr*.anno.vcf.gz*
            2. Merged chromosomes :: gs://dinglab/isb-cgc/tcga/germline/production/merge.ExAConly/all/annotated/by_analysisID/PCA.merge.ExAConly.analysisID.anno.vcf.gz*
    b. All samples called, by cancer type, unannotated :: gs://dinglab/isb-cgc/tcga/germline/production/merge.ExAConly/by_cohort/${cancer}.merge.ExAConly.analysisID.vcf.gz*

    c. Samples called, annotated, whitelisted
        i. ***By analysisID ::*** gs://dinglab/isb-cgc/tcga/germline/production/merge.ExAConly/all/annotated.by_chr/whitelist.by_chr/by_analysisID/PCA.merge.ExAConly.analysisID.chr*.anno.whitelist.vcf.gz*
        ii. ***By TCGA barcode ::***
            1. By chromosome :: gs://dinglab/isb-cgc/tcga/germline/production/merge.ExAConly/all/annotated.by_chr/whitelist.by_chr/by_tcgaBarcode/PCA.merge.ExAConly.tcgaBarcode.chr*.anno.whitelist.vcf.gz*
            2. Merged chromosomes :: gs://dinglab/isb-cgc/tcga/germline/production/merge.ExAConly/all/annotated/by_tcgaBarcode/PCA.merge.ExAConly.tcgaBarcode.anno.whitelist.vcf.gz*

b. Coverage data
    i. By coding interval
        1. Data: gs://dinglab/isb-cgc/tcga/germline/production/coverage.Encode/coverage.Encode.mpileup.q1_Q13.<analysisID>
        Format: chr, start, stop, <ignore>, total number of bases in interval, <ignore>
        Bed file: See item (1)(h)(ii) above
    ii. miRNA intervals
        1. Data: gs://dinglab/isb-cgc/tcga/germline/production/coverage.regions.mirBase/
        Bed file:
        2. gs://dinglab/isb-cgc/tcga/germline/production/coverage.regions.mirBase.mature/
        Bed file:

12. **gs://dinglab/isb-cgc/tcga/manifests/**
    a. 201710/CGHub_legacyGDC_DNA_bams.csv.gz
       TCGA manifest of
    b. gs://dinglab/isb-cgc/tcga/manifests/02jun2016/*tsv
       TCGA manifests on the ISB cloud (Listing of samples with available WXS and WGS by cancer)