Outline

- We uniformly annotate PCAWG variants and provide this as a resource to the wider community.
    - This resource will serve as a comprehensive non-coding annotation compendium of cancer variants.

    - We utilize this resource to further quantify mutational burdening of various genomic elements for different cancer subtypes
        - LOF/non-synonymous variants (gene-centric analysis)
        - Noncoding variants (gene centric analysis)

    - We also annotate and quantify the enrichment of large structural variations identified in the PCAWG project
        - Germline vs somatic enrichment
        - Engulf vs partial overlap

- Alteration in transcriptional regulation is often associated with tumorigenesis, thus we closely inspect the mutational burdening of various transcription factor and their target genes.
    - SNV induced Gain/loss of TF motif events among different TFs in various cancer subtypes
        - Enrichment of SNV compared to genomic background

    - Closer inspection of the affected target genes

    - Downstream gene expression analysis

- In addition to annotation, we also evaluate the impact of PCAWG variants including SNVs, INDELs and large SVs.
    - We can categorize each SNV into three distinct class based on their impact score: a) low impact passenger variants, b) medium impact passenger and c) high impact putative driver
        - We provide breakdown of these categories of variants in different cancer subtypes. This include coding and non-coding variants.

    - We observe multimodal functional impact distribution with peak in medium impact regime suggesting significant presence of medium impact passengers
        - Certain cancer types are highly enriched among these medium impact passengers
        - These medium impact passengers are prevalent among essential genes

- SV impact score distribution suggest presence of high impact deletion and duplications in sarcoma and glioma cohorts
  - Gene level enrichment analysis

- We also delineated the underlying signatures of various SNVs
  - We observe distinct signature composition for high impact, medium impact and low impact SNVs
    - This highlight role of distinct mutation processes in emergence of different categories of variants
  - In addition, we also calculated the mutation spectra of SNVs leading to gain or loss of motif events in highly mutated transcription factor.

- Furthermore, we also integrated subclonal information of each variant to infer the evolution of impactful variants during cancer progression
  - High impact mutations are more prevalent in early subclone compared to later
    - As expected, this observation is more pronounced for high impact LOF variants in TSG
  - Clonal mutational burden analysis
  - VAF distribution for LOF variants

- We explored presence of weak positive/negative selection among passenger variants
  - Description of the different categories of SNVs
    - Deleterious passengers, latent driver and mini driver
  - Extreme VAF & selection among passengers
    - Co-mutation, impact score and VAF to identify candidate variants
  - Somatic and germline mutation load and prediction of cancer onset age
  - Survival analysis and latent driver/deleterious passenger
    - CLL and Kidney-RCC results

Introduction

Cancer progression is an evolutionary process during which thousands of somatic variants are accumulated within an individual. In the classic view of cancer progression, a handful of driver variants are thought to give a positive selection advantage to the cancer cell. In contrast, majority of these variants, often labeled as passengers, confer no selective advantage to the cancer cell and are considered to occur neutrally. This canonical dichotomy is often useful, but it is also imprecise. A more nuanced view admits that some passenger variants may impact tumor cell biology along a range of dimensions and weakly affect tumor cell fitness for better or for worse. However, maximum number of passenger variants occupy non-coding regions of the genome and thus make it very challenging to interpret their functional consequences.

Previous studies have extensively focused on characterizing variants occupying coding regions of various cancer genomes. However, the exhaustive PCAWG variant dataset, which comprises pan-cancer variant calls from ~2700 uniformly processed whole cancer genomes, gives us an unparalleled opportunity to investigate the overall functional burdening of different non-coding genomic elements. Given that the majority of cancer variants lie in non-coding regions, this variant dataset serves as a substantially more informative resource than the many existing datasets focused on exomes. In addition, it also contains a full spectrum of variants, including copy number variants (CNVs) and large structural variants (SVs) in addition to SNVs and INDELs.

In this work, our purpose is two-fold: First, we build on and apply existing tools to annotate and score the predicted functional impact of each variant, including SNVs, INDELs and SVs in the pan-cancer dataset. This systematic annotation effort generates a comprehensive annotation compendium of PCAWG variants, which can serve as a useful resource for the wider cancer genomics community. Furthermore, we leverage this resource to further characterize variant induced overall functional burdening of different genomic elements for various cancer cohorts. Subsequently, we test the underlying hypothesis of the classic view that all passenger variants are neutral and they don't play any role in cancer progression. We find several forms of evidences indicating weak (positive and negative) selection of passenger variants, particularly those we predict to be impactful. We find that the distribution of impact scores among passenger variants, their mutational signatures, co-mutation frequencies, ability to predict age of cancer onset (in the case of germline variants), patient survival, and inferred evolutionary timing contradict the classical view that all passenger variants are neutral.