PertInInt

Goals

- Identify genes with functional roles in cancer
  - Reveal similarities between different cancer types
  - Reveal how differences between populations of tumors cells may impact treatments
- Uncover genes mechanisms of action — Understand cellular processes in tumors

Methods

- Classic framework — Drawback
  - Don't detect less frequent mutational event
  - Don't allow to understand the mechanisms
  - Consider somatic mutation only alter a single type of functionality
  - "Black box " machine learning approach
  - Interaction sites only identified for human specific genes
- Cell Systems integrative framework
  - Advantages
    - Avoid time-prohibitive permutation-based significance tests — Feasibility to simultaneously consider multiple measure
    - Pin pointing genes with somatic mutation across tumors — Different mutations within a gene can have unequals impacts : need of a "sub-gene" analyse
    - Consider somatic mutation alter a broad range of protein functionalities — Types of functionalities
      - Evolutionary conservation
      - 3D structure
      - Domains
      - Post-translation modification
    - Interaction sites identified for 63% of human genes
  - Tracks

Run PertInInt on the pan-cancer dataset when restricted to each of these track types in turn, to identify which is useful

- **Interaction tracks** — Interaction tracks correspond to portions of a protein that are inferred to interact with ligands
  1. Per-position weights reflect the observed residue-to-ligand proximities, computed as the fraction of atoms in the amino acid residue found within 4.0A˚ of the ligand.
  2. Each position within an InteracDome domain is associated with a "binding frequency" between 0 and 1 that corresponds to the fraction of the time residues in this position were found to be in contact with the ligand of interest when analyzing co-complex structures
  3. We also consider additional tracks encoding multiple instances of the same domain family in a protein; these "aggregate" tracks span noncontiguous intervals that correspond to the locations of individual domain instances, with track positions weighted according to the binding frequencies at corresponding domain match states as described above.

  → Numerous tracks per gene
- **Domain tracks**
  - Weights are 1 for amino acid (domain of interest) positions and 0 elsewhere
  - Domain tracks span the length of the protein
- **Conservation tracks** — Weights measures the protein sequence conservation across vertebrate homologs — Weights are obtained by multiplying the fraction of non-gap residues in the column by the Jensen-Shannon divergence (JSD) between those non-gap residues and a Blosum 62 background amino acid distribution

  → One track per gene
- **Natural variation tracks** — Single entry per gene, weights correspond to natural variation estimation within healthy genome

  - Scores
    - For each track, we consider the somatic mutations observed across tumor samples that fall within positions of that track and compute a per-track score as the sum of the per-track weights of the positions that each of the mutations fall into
    - Reference score (expected by chance)
      - Using mean and standard deviation — Shuffle the mutations across the positions of the track, and use the mean and standard deviation computed from these permutations to compute a Z score
      - Analytically (7x faster) — For each protein, we next combine the information from each of its tracks. Because tracks can overlap along the length of the protein sequence, and the somatic mutations that fall in each of them can also overlap, these tracks cannot be treated independently. Instead, for the background model we derive an approach to compute the covariance be- tween tracks analytically and then use this covariance matrix to estimate a combined score

      Compared

    - Reference score — Based on the Cancer Gene Census (CGC)

Results

- Multiple sources of informations
  - Comparing to CGC
    - Each track: recapitulate known CGC genes with varying degrees
    - Interaction track: identify the largest number of knownCGC genes
  - Performances: three sources > two sources > on source
  - Importance of between track covariance
  - Cancer driving genes identified by each four tracks by turn: less than 10%
  - Low mutated genes harbor mutations that preferentially alter functional sites
- Mutations distributed across interaction interfaces — Are mutations within a small number of interaction sites or across several interaction sites
  - High entropy: mutations spread aross many interaction sites
  - Low entropy: mutations patterns can be uncovered by mutation "hotspots" methods

  → PertInInt highly ranks several hotspots ut also find others genes with perturbed interactions
  - Analysis reveal top-ranked genes include both oncogenes and tumor suppressor genes — Oncogenes tumors enrichment : 2.36 greater than TSGs enrichment