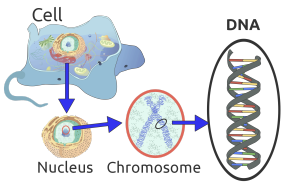
PertInInt.

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/350cbdcc-0de0-46a0-8857-93e71a684508/Kobren__al_-_PertInInt_(2020).pdf

▼ *Table of content*

Genetic concepts.



Cell structure.

Each **chromosome** is composed of DNA.

DNA contain a nucleotid sequence (red, yellow, blue and green on the picture).

Each **nucleotid** triplet is a codon.

Functioning.

A **codon** code for an amino acid.

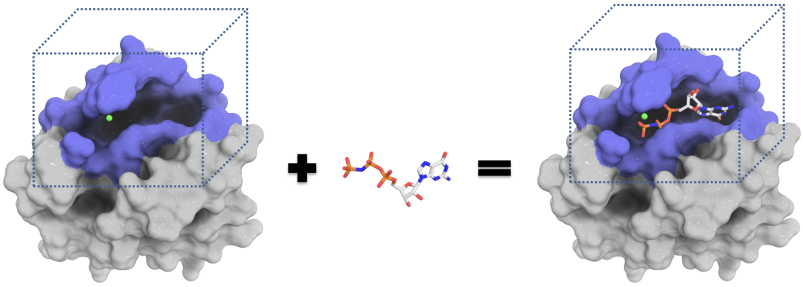
1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine †	TCT	(Ser/S) Serine †	TAT	(Tyr/Y) Tyrosine †	TGT	(Cys/C) Cysteine †	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine †	TCA		TAA	Stop (*Ochre*) *[note 2]	TGA	Stop (*Opal*) *[note 2]	A
	TTG →		TCG		TAG	Stop (*Amber*) *[note 2]	TGG	(Trp/W) Tryptophan †	G
C	CTT		CCT	(Pro/P) Proline †	CAT	(His/H) Histidine ‡	CGT	(Arg/R) Arginine ‡	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine †	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine †	ACT	(Thr/T) Threonine †	AAT	(Asn/N) Asparagine †	AGT	(Ser/S) Serine †	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine ‡	AGA	(Arg/R) Arginine ‡	A
	ATG →	(Met/M) Methionine †	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine †	GCT	(Ala/A) Alanine †	GAT	(Asp/D) Aspartic acid ↓	GGT	(Gly/G) Glycine †	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid ↓	GGA		A
	GTG →		GCG		GAG		GGG		G

Several codons with corresponding amino acid.

Several **amino acids** forms a **protein**.

A **mutation** is an anormal change of a nucleotid, that may impact the amino acid sequence and therefore the protein shape.

Each healthy human possesses a lot of mutations.



A protein (grey) with functional site (purple) and ligand. These two elements form a complexe.

Cancer origins.

Heatlhy functioning.

Replication is when the cell divide itself to produce two new cells.

Apoptosis is when the cell die by itself, to allow tissus regeneration.

Tumor appearance.

When cells division occur without control, it forms a tumor that may cause harm to surrounding tissues.

PertInInt study consider two gene types :

- **Oncogenes** : responsible of anormal cell replication leading to the tumor (replication or apoptosis problem).
- **Tumor Suppressor Gene** (TSG) : control the cell replication in case of tumor.

Both have an enrichment of somatic mutations within interaction interfaces.

They also take into account **Putative Cancer Genes** : sequence of DNA that is believed to be a gene, can share sequence similarities to already characterized genes.

Somatic mutations.

Drivers are somatic mutations with functional roles in cancer (proportion part of somatic mutations).

Passengers are neutral mutations.

Differentiating these 2 types of mutations is still computationally difficult. Plus, different positions within genes = different molecular functionalities —> methods using **subgenes** to study specific mutation sites.

PertInInt.

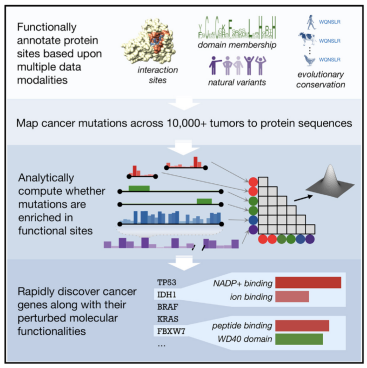
Goals.

Identify genes (somatic mutations) potentially implied in tumor developpement.

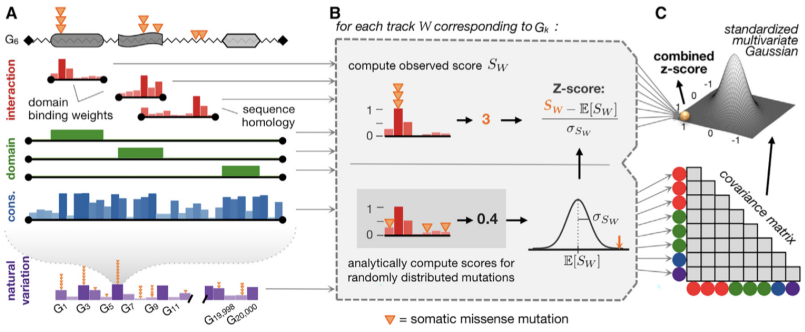
Understand their predictions, in opposition to classical ML models (black box).

Methods

- Creating **tracks**, corresponding to various aspects of a protein.
- Each track associate with nucleotides a **weight** (0 to 1) according to its importance with respect to the corresponding aspect.
- Therefore, some genes with fewer mutation but associated with major protein functionality may be detected by *PertInInt*.
- Data:
  - 10 037 tumor samples.
  - 33 tumor types.
  - Set genes form *Cancer Gene Census* (CGC).



PertInInt's operating principle.



Four tracks' integration.

Tracks definition.

- **Interaction**:
  - Per-position weights reflect the observed residue-to-ligand proximities, computed as the fraction of atoms in the amino acid residue found within 4.0Å (40^{-10}meters) of the ligand.
  - Higher positional weights = positions most likely to take part in interaction with a ligand (based on known protein positions with ligand-binding potential).
  - 63% genes have per-site information about interactions.

- **Domaine**:
  - Domain tracks span the length of the protein, and positions within and outside of the domain instance are respectively assigned weights of 1 and 0.
  - 90% genes have per-site information about domains.
- **Conservation**:
  - Length of protein sequence, mesures conservation across vertebrate homologs.
  - Weights are obtained 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.
  - Higher weights = positions under more evolutionary constraint.
  - All genes have per-site conservation value.
- **Natural variation**:
  - Background mutation rate, estimated from the number of variants across healthy populations.
  - All genes have background gene-level mutation rates.

Where tracks overlap, the covariance is computed (combined score estimated by covariance matrix). If not, leads to worse (poor) cancer-revelant genes detection.

Score computation.

- **Mutation** score:

$$S_W = \sum_{i=1}^{n} f_i z_i$$

  - n is the number of mutation.
  - z is the weight on the position where the mutation appears.
  - f is the proportion of sequencing reads that contain the mutation.
- Analytical **Z-score**:

"

- **Background** mutational model:

$$\lambda_j = \sum_{d \in 1,2,3} (B_{jd} \times \sum_{U \in \{A,T,G,C\}} M_{jdu})$$

  - Where

$$B_{jd} = \begin{cases} 1 & \text{if the } d\text{th nucleotide in the codon at position } p_j \text{is A or T.} \\ b & \text{otherwise, where } b \text{ is the relative frequency of a C/G mutation} \\ & \text{in the pan-cancer dataset as compared to a A/T mutation.} \end{cases}$$

  - And

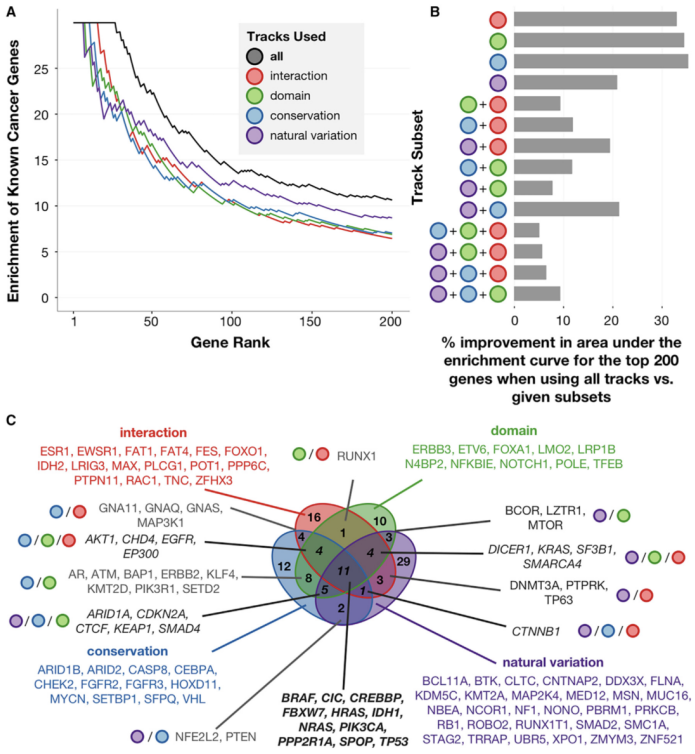
$$M_{jdu} = \begin{cases} 1 & \text{if changing the } d\text{th nucleotide in the codon at position } p_j \text{ to } u \text{ results} \\ & \text{in a missense mutation.} \\ 0 & \text{otherwise.} \end{cases}$$

- **Covariance** between tracks.

- We obtain per-gene score combining the four tracks.

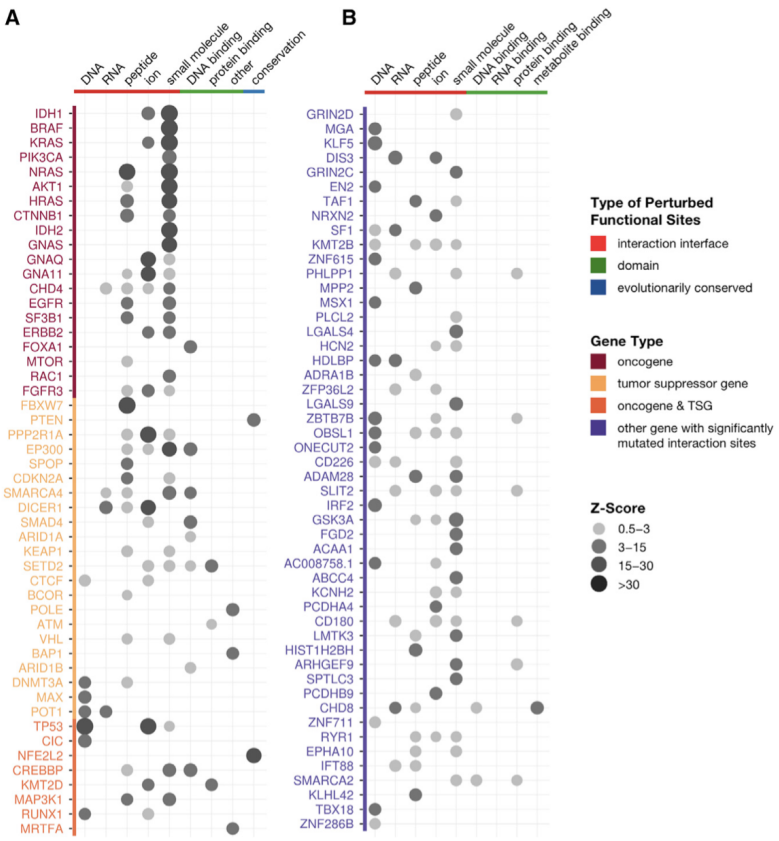
Results.

- Using all tracks together lead to the best results :
  - Interaction.
  - Domain.
  - Evolutionary conservation.
  - Whole-gene mutation frequency (natural variation).



Interaction of the four tracks.

- All 4 tracks types identify less than 10% of CGC genes.
- Low mutated genes harbor mutations that preferentially alter functional sites.
- Analysis reveal top-ranked genes include both oncogenes and tumor suppressor genes.
- **Oncogenes tumors** enrichment : **2.36** greater than **TSG**s enrichment.
- *PertInInt* can process the pan-cancer mutational data while considering multiple sources of data about protein functionality in 10 min on a single core of a standard desktop, compared by several minutes or days for other methods.



Z-scores for oncogenes, TSGs & putative drivers genes.

Discussion.

QuaDMutNetEx: a method for detecting cancer driver genes with low mutation frequency - BMC Bioinformatics

Background Cancer is caused by genetic mutations, but not all somatic mutations in human DNA drive the emergence or growth of cancers. While many frequently-mutated cancer driver genes have already been identified and are being utilized for diagnostic, prognostic, or therapeutic purposes, identifying driver genes that harbor mutations occurring with low

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3449-2

To Do.

👩‍💻 Source Code