Analyzing genome sequencing data to identify cancer drivers

Eric Minwei Liu
Khurana lab
June 19th, 2019
Outline

• Background

• Examples of non-coding cancer driver projects
  – **CNCDriver**: new algorithm to detect non-coding cancer drivers
    • CTCF/cohesin insulators as a class of functional element
    • Functional validation of insulator mutations
  – **CNCDatabase**: non-coding cancer drivers database

• Future perspective
Cancer arises from DNA mutations in cells

Driver and passenger mutations

• **Driver mutations**
  – Confer growth advantage to cancer cells
  – Selectively accumulated in the process of tumor development

• **Passenger mutations**
  – Somatic mutations with no functional consequences
  – Mutations do not provide growth advantage to cancer cells
Cis-regulatory elements control gene expression

- Promoter / enhancer / insulator

Schematic data analysis workflow

WGS on cancer & non-cancer DNAs (90-100 Gb~ sequences)

Alignment
- BWA
- BWAmem
- Novoalign
- SNAP
- GEM

[FASTQ files]

Remove of PCR duplication (2-10%) and other filtering

Somatic mutation call
- SNVs
  - Mutect
  - Strelka
  - VarScan2
  - UnifiedGenotyper
  - SomaticSniper
  - SMuFin
- Short indels
  - Strelka
  - Pindel
  - Dindel
  - Indelocator
  - SOAPindel
- SVs
  - dRanger
  - BreakDancer
  - BreakPointer
  - GRAFT
  - DELLY

[FASTQ files]

Copy number
- ABSOLUTE
- PICNIC
- HMMCOPY
- DNAcopy
- VarScan2

[VCF files]

Molecular annotation and interpretation

Driver gene discovery
- MuSiC
- MutSig
- OncodriveCLUST
- OncodriveFM

Mutational signature
- Non-negative matrix factorization (NMF) and principal component analysis (PCA)

Problem of mutational heterogeneity

uniform model

heterogeneous model

Background mutation rate correlates with multiple factors

- Replication timing
- Gene expression

![Graph showing replication timing and mutation rate correlation](image)


- Nucleotide excision repair (NER)

![Diagram showing nucleotide excision repair](image)


- Sequence context (mutation signature)

![Signature 7 graph](image)


- Nucleosome region

![Diagram showing nucleosome region](image)

**TERT** promoter frequently mutated in multiple cancer types

<table>
<thead>
<tr>
<th>Tumor Type</th>
<th>Mutation Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skin tumor</td>
<td></td>
</tr>
<tr>
<td>Basal cell carcinoma</td>
<td>132/278 (47.4)</td>
</tr>
<tr>
<td>Squamous cell carcinoma</td>
<td>75/125 (60.0)</td>
</tr>
<tr>
<td>Merkel cell carcinoma</td>
<td>5/49 (10.2)</td>
</tr>
<tr>
<td>Pleomorphic dermal sarcoma</td>
<td>26/34 (76.0)</td>
</tr>
<tr>
<td>Atypical fibroxanthoma</td>
<td>25/27 (93.0)</td>
</tr>
<tr>
<td>Malignant melanoma</td>
<td></td>
</tr>
<tr>
<td>Cutaneous melanoma</td>
<td>564/1287 (43.8)</td>
</tr>
<tr>
<td>Other types of melanoma</td>
<td>165/505 (32.7)</td>
</tr>
<tr>
<td>Brain tumor</td>
<td></td>
</tr>
<tr>
<td>Glioma (low-grade)</td>
<td>929/2580 (36.0)</td>
</tr>
<tr>
<td>Glioma (high-grade)</td>
<td>2171/3085 (70.4)</td>
</tr>
<tr>
<td>Meningioma</td>
<td>25/337 (7.4)</td>
</tr>
<tr>
<td>Medulloblastoma</td>
<td>36/182 (19.8)</td>
</tr>
<tr>
<td>Endocrine tumor</td>
<td></td>
</tr>
<tr>
<td>Thyroid cancer</td>
<td></td>
</tr>
<tr>
<td>Papillary thyroid carcinoma</td>
<td>583/5380 (11.0)</td>
</tr>
<tr>
<td>Follicular thyroid carcinoma</td>
<td>59/346 (17.1)</td>
</tr>
<tr>
<td>Anaplastic thyroid carcinoma</td>
<td>93/237 (39.2)</td>
</tr>
<tr>
<td>Hurthle cell carcinoma</td>
<td>8/61 (13.1)</td>
</tr>
<tr>
<td>Atypical follicular thyroid adenoma</td>
<td>3/18 (16.7)</td>
</tr>
<tr>
<td>Differentiated thyroid carcinoma</td>
<td>41/339 (12.1)</td>
</tr>
<tr>
<td>Poorly differentiated thyroid carcinoma</td>
<td>73/170 (42.9)</td>
</tr>
<tr>
<td>Adrenocortical carcinoma</td>
<td>4/98 (4.1)</td>
</tr>
<tr>
<td>Gynecological tumor</td>
<td></td>
</tr>
<tr>
<td>Ovarian clear cell carcinoma</td>
<td>48/301 (15.9)</td>
</tr>
<tr>
<td>Ovarian low grade serous</td>
<td>2/41 (4.9)</td>
</tr>
<tr>
<td>Endometrial carcinoma</td>
<td>5/78 (6.6)</td>
</tr>
<tr>
<td>Squamous cell carcinoma of the cervix</td>
<td>33/335 (9.9)</td>
</tr>
</tbody>
</table>

**TERT** promoter mutation create *de novo* ETS binding site

Liu et al. *Genes* (2016)  
Pan-cancer analysis of 1962 whole-genome sequencing samples across 21 cancer types

1962 whole-genome cancer samples

- 1446 samples
  - Alexandrov et al
  - Fredriksson et al
  - Perera et al
- 452 samples
  - ICGC
- 64 samples
  - Baca et al
  - Berger et al

Fredriksson et al. Nature Genetics, (2014)
Baca et al. Cell (2013)
CTCF/cohesin insulators as a class of functional element

- CTCF and cohesin have roles in modulating chromatin structure and gene expression

cohesin ChIA-PET CTCF ChIA-PET
cohesin pull-down CTCF pull-down
Chromatin loop disruption can activate oncogene in T cell acute lymphoblastic leukemia (T-ALL)

Hnisz et al. Science 351, 1454-1458 (2016)

LMO2 has 2 fold over-expression when boundary site was deleted by CRISPR/Cas9
Working model

- CTCF motif loss mutations disrupt CTCF binding at insulator and DNA loop.
Working model

- CTCF motif loss mutations disrupt CTCF binding at insulator and DNA loop.
Motivation

• Identify which CTCF/cohesion insulators are putative cancer drivers

• Identify CTCF-CTCF loop disruption that associates with aberrant gene expression
CTCF/cohesin insulator annotations from ChIA-PET data

Consensus CTCF/cohesin insulators from intersection of at least 5 out of 7 cohesin ChIA-PET or CTCF ChIA-PET data (GM12878, K562, Jurkat, MCF-7 and HeLa-S3)

~16k CTCF/cohesin insulators
Median length ~ 2 Kb

15 out of 21 cancer types have significant enrichment of CTCF motifs predicted to be disrupted in the genome

A

CTCF motif-break

For example, C to A mutation if $M_{13} < M_{33}$, motif-breaking

<table>
<thead>
<tr>
<th>PWM</th>
<th>Pos1</th>
<th>Pos2</th>
<th>Pos3</th>
<th>Pos4</th>
<th>Pos5</th>
<th>Pos6</th>
<th>Pos7</th>
<th>Pos8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>M_{11}</td>
<td>M_{12}</td>
<td>M_{13}</td>
<td>M_{14}</td>
<td>M_{15}</td>
<td>M_{16}</td>
<td>M_{17}</td>
<td>M_{18}</td>
</tr>
<tr>
<td>G</td>
<td>M_{11}</td>
<td>M_{22}</td>
<td>M_{23}</td>
<td>M_{24}</td>
<td>M_{25}</td>
<td>M_{26}</td>
<td>M_{27}</td>
<td>M_{28}</td>
</tr>
<tr>
<td>C</td>
<td>M_{11}</td>
<td>M_{32}</td>
<td>M_{33}</td>
<td>M_{34}</td>
<td>M_{35}</td>
<td>M_{36}</td>
<td>M_{37}</td>
<td>M_{38}</td>
</tr>
<tr>
<td>T</td>
<td>M_{11}</td>
<td>M_{42}</td>
<td>M_{43}</td>
<td>M_{44}</td>
<td>M_{45}</td>
<td>M_{46}</td>
<td>M_{47}</td>
<td>M_{48}</td>
</tr>
</tbody>
</table>

Fu et al. Genome Biology (2014)
Higher rate of CTCF motif-break mutations are attributed as neutral mutations

- Although CTCF motif-break mutations have higher functional impact and are enriched, they are neutral mutations due to association with signatures of mutational processes.

Red: CTCF motif-break mutations
Grey: CTCF motif-preserving mutations
Dot line are expected mutation rate from random simulation
CNCDriver detects regions with significantly more functional bias than expected at random.

1. Compute CNCDriver score ($S_j$) for insulator $j$

$$S_j = \sum_{i=1}^{n} W_i \times FS_i$$

- $S_j$ = CNCDriver score
- $j$ = insulator $j$ in the genome
- $n$ = total number of mutated positions
- $W_i$ = fraction of mutated samples at position $i$
- $FS_i$ = functional impact score at position $i$

2. Build a null background for $S_j$

- $S_{j1}$
- $S_{j2}$
- $S_{j3}$
- $\vdots$
- $S_{jN}$

- Mutational covariates
- Element type
- Trinucleotide sequence probability
- Replication timing
- Ratio of mutations predicted to disrupt CTCF motifs or not

3. $P$-value ($P_j$) estimation

$$P_j = \frac{1 + \text{count}(S_{jk} \geq S_j)}{N + 1}$$

- $S_{jk}$ = simulated CNCDriver score computed by sampling $n$ mutations accounting for covariates in insulator $j$. ($k \in \{1, N\}; N=10^5$)
Functional impact score framework in modified FunSeq2

\[
F_{I_{\text{noncoding}}} = w_1A + w_2B + w_3C + w_4D + w_5E + w_6F + w_7G
\]

\[
w_d = 1 + P_d \log_2 P_d + (1 - P_d) \log_2 (1 - P_d) \quad [1 - (\text{shannon entropy})]
\]

\[
P_d = \frac{\text{number of polymorphisms with score} \geq \text{feature d}}{\text{total number of polymorphisms in 1000 genomes}}
\]

<table>
<thead>
<tr>
<th>variables</th>
<th>features</th>
<th>details</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>functional annotations</td>
<td>Functional annotation from ENCODE</td>
</tr>
<tr>
<td>B</td>
<td>human population-level conservation</td>
<td>In sensitive/ultra-sensitive regions</td>
</tr>
<tr>
<td>C</td>
<td>evolutionary conservation</td>
<td>In ultra-conserved regions / Genomic Evolutionary Rate Profiling (GERP) score &gt;2</td>
</tr>
<tr>
<td>D</td>
<td>in highly occupied target (HOT) regions</td>
<td>TF highly occupied regions</td>
</tr>
<tr>
<td>E</td>
<td>in regulatory elements associated with genes</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>CTCF motif break</td>
<td>PWM changes</td>
</tr>
</tbody>
</table>

Fu et al, Khurana et al (2013)
CNCDriver results in melanoma (N=219)

- Melanoma (SKCM-US:36 and MELA-AU: 183)

**Coding sequence**

**promoter**

**enhancer**

**lincRNA**

OncodriveFML
21 significantly mutated CTCF insulator sites in total

- 21 significantly mutated CTCF insulator sites in total (pan-cancer and single cancer type)
- 19 CTCF insulators sites with gene expression data.
Example of mutations in the insulator driver candidate

1. neighboring insulator with convergent CTCF motif orientations.
2. Within 360 Kb distance (75th percentile of constitutive CTCF-CTCF loop length).
3. Within the same TAD.

Criteria to determine affected chromatin neighborhood
Insulator mutations are associated with \textbf{TGFBI} up-regulation in the patients data

\textbf{Differential expression of genes in hotspot mutated samples}

\begin{itemize}
  \item \textit{CYP2S1} \(q = 0.008\)
  \item \textit{TGFBI} \(q = 0.006\)
\end{itemize}

\begin{tabular}{ll}
  WT & MUT  \\
  N=69 & N=11  \\
  WT & MUT  \\
  N=69 & N=11
\end{tabular}

3C assay and CTCF ChIP-seq support A-B loop existence in human A375 melanoma cells

CTCF binding from ChIP-seq

Loop validation using 3C

Elissa Wong
Ping Chi

Boaz Aronson
Effie Apostolou
Prior studies suggest ELK4 can also co-localize with CTCF at loop anchors to affect long-range chromatin interactions.

Design CRISPR assay to interrogate functions of SNV4 and SNV8 in the human A375 melanoma cells

- We designed 4 sgRNA sequence targeting SNV4 and SNV8.
Deep sequencing to determine indel spectrum

• Most indel size is around -10bp.
CRISPR perturbations support functional roles of SNV4 and SNV8

- SNV4 and SNV8 perturbations by CRISPR increase cell growth rate 40% to 50%
Summary

• CTCF/cohesin insulators are a potential novel class of cancer drivers

• Predicted candidate interfere with TGF-β pathway.

• CNCDriver also could be used to predict cancer drivers for CDS, promoter, enhancer, and lincRNA.
Curated cancer gene list from literatures

- There are many information about coding cancer drivers
- No database collecting information for non-coding cancer driver yet
Cornell Non-coding cancer driver database (CNCDatabase)

1. Pubmed / Text mining
2. User submitted data

Add data to database

Data curation queue

SQL query / Web API

Data / Visualization

Front-end
User interface

Back-end
Database
Interactive web user interface

Gene overview (provided by HGNC)

- Symbol: WDR74
- Approved name: WDR74
- HGNC ID: 26609
- Ensembl ID: ENSG00000133316
- NCBI Gene ID: 64663
- Local type: protein_coding
- Description: WD repeat domain 74 [Source: HGNC Symbol; Accession: 26609]
Summary of number of genes associated with non-coding drivers in 31 cancer types

Computational prediction
- 657 genes from pancancer analysis
- 587 genes from single cancer analysis
- On average 12 genes associated with non-coding cancer drivers

B) Gene expression association using RNA-seq

C) Experimental validation

- 10 genes have gene expression from RNA-seq
- 19 genes are from experimental validation
TERT, WDR74, PLEKHS1 and CCDC107 are top 4 most predicted promoter candidates
Summary

• CNCDatabase will be a helpful resource to query the evidences that support non-coding driver mutations in cancer.

• This database resource can also help to prioritize the functional validation list.
Future perspective

• More non-coding drivers are yet to be revealed.
  – Analyze more whole genome samples
  – Integrate WGS mutations with other types of data

• Large-scale functional screens to study non-coding variants.
Acknowledgements

- Ekta Khurana
- Alexander Martinez Fundichely
- Tawny Cuykendall
- Priyanka Dhingra
- Tuan Trieu
- Fanying Tang
- Andre Neil Forbes
- Matthew Mackay
- Erica Duo Xu

- Neville Sanjana (NYGC)
- Bianca Diaz (NYGC)
- Effie Apostolou
- Boaz Aronson
- Ping Chi
- Elissa Wong
- Steven Lipkin
- Chason Lee

- Christina Leslie
- Olivier Elemento
CNCDriver CDS results
CNCDriver promoter results
CNCDriver enhancer results
CNCDriver lincRNA results