LineUp

Identifying Deleterious Mutations Using Protein Domain Alignment

Brady Bernard
Shmulevich Group
Normalized mutation frequencies
Twenty tumor types

Mutations/Protein Length

Mutated Gene Rank

TP53
KRAS
PTEN
Normalized mutation frequencies
Twenty tumor types

Most genes are mutated at low frequency
Which mutations are deleterious?
Example protein mutation spectrum
Pan-cancer mutation spectrum
Pan-cancer Ras domain mutation spectra

NRAS

Stomach

Pan-cancer

HRAS

Pan-cancer

KRAS

Pan-cancer

log_2 scaled counts per position
Ras domain structure alignment

- KRAS
- NRAS
- HRAS
- GDP
- Hotspots
Typical pan-cancer mutation spectrum

<table>
<thead>
<tr>
<th>Tumor Type</th>
<th>MutSig Rank</th>
<th>Samples (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLCA</td>
<td>9638</td>
<td>1</td>
</tr>
<tr>
<td>BRCA</td>
<td>952</td>
<td>0</td>
</tr>
<tr>
<td>GBM</td>
<td>14085</td>
<td>0</td>
</tr>
<tr>
<td>KIRC</td>
<td>2051</td>
<td>1</td>
</tr>
<tr>
<td>LUAD</td>
<td>925</td>
<td>14</td>
</tr>
<tr>
<td>LUSC</td>
<td>9703</td>
<td>2</td>
</tr>
<tr>
<td>OV</td>
<td>3421</td>
<td>2</td>
</tr>
<tr>
<td>STAD</td>
<td>1118</td>
<td>8</td>
</tr>
<tr>
<td>UCEC</td>
<td>2211</td>
<td>7</td>
</tr>
</tbody>
</table>
LineUp Approach

- Align sequences from all matching domains across all tumor types
- Evaluate missense mutation frequencies
- Applies to ~40% of missense mutations
Mutations per position for two hundred aligned Homeobox domains across twenty tumor types
Identifying outliers: Modified Z-score

MAD (Median of Absolute Deviations about the median):
\[ MAD = \text{median}\{|x_i - \bar{x}|\} \]

Modified Z-score \[ M_i = \frac{0.6745(x_i - \bar{x})}{MAD} \]
Identifying outliers: Modified Z-score

\[ M_4 = 8.3 \]

MAD (Median of Absolute Deviations about the median):
\[ MAD = \text{median}\{|x_i - \bar{x}|\} \]

Modified Z-score
\[ M_i = \frac{0.6745(x_i - \bar{x})}{MAD} \]
All positions within all domains
High Modified Z-score and counts for Ras domain

Position 12 in Ras domain
KRAS, NRAS, HRAS
Multiple other Ras genes mutated at the same position

Position 12 in Ras domain
KRAS, NRAS, HRAS
RAP1B, RAB7L1, RALB, RIT2, RAB25, RHOB, NKIRAS1, RAP2B, RAB3C, RAB5C
Investigating genes with low mutation frequency
Investigating genes with low mutation frequency
Background expectation:
Comparing TCGA to 1000 Genomes
Outlier position within Homeobox domain

Homeobox Position 4
48 Genes:
ISX, HDX, HOXA1, ONECUT2, LHX2, POU4F3, VAX1, ZFHX3, NKX3-1, CDX4, NOBOX, CUX1, SATB2, PBX2, DPRX, MSX2, LHX1, HOXC13, DUXA, DUXA, HOXC9, HMX3, HOXB6, EVX2, ISL1, HOXB2, LMX1A, ZHX1, HOXA10, DBX2, LHX9, SATB1, HOXD4, LMX1B, BARX2, MEOX2, ALX4, HMBOX1, HOXD9, HOXA4, LHX4, LBX1, DLX2, HOXB13, PAX7, EVX1, HOXA5, HOXC8
Structural interpretation
Summary and Next Steps

• We have comprehensively evaluated mutations at all positions within all domains to identify low frequency but likely deleterious mutations

• Hotspots outside of domains and mutations that broadly disrupt structure and function not addressed
  – Integration with other methods is essential

• Functional validation of low frequency events in such data sets remains challenging

• As cancer and normal genome data size increases, more robust normalization per position per domain can be achieved
Acknowledgements

Ilya Shmulevich
Tae Denwongkun
Theo Knijnenburg

Vesteinn Thorsson  Sheila Reynolds
Kalle Leinonen  Lisa Iype
Richard Kreisberg  Natalie Tasman
Hector Rovira
2bnh: alpha-beta horseshoe
1hv9: left-handed beta helix
1m30: SH3-like barrel
Example approaches and methods

- Statistical assessment of multiple samples (MutSig, MuSiC, …)
  
  **Mutational heterogeneity in cancer and the search for new cancer-associated genes**
  
  Michael S. Lawrence1,*, Petar Stojanov1,†, Paz Png1, Scott L. Carter1, Chip Stewart1, Craig H. Merlent2,3,4,5,6, Vidyamurthy Dicks1,3,4,5,6, Libhua Zou1, Alex H. Ramos1,7, Taisong Lu1, Carrie Sougnez1, Lauren Ambrogi1, Elizabeth Nicola1, Douglas Voel1, Michael Noble1, Daniel DiCara1, Peter Nowak1, Peter Will1, Paul V. Lemkin1, Peter Donahoe1, Michael J. Sklar1,4,5,6

- Sequence conservation (MutationAssessor, SIFT, …)
  
  **MuSiC: Identifying mutational significance in cancer genomes**
  
  Nathan D. Dees,1,4 Quynuan Zhang,1,4 Cyriac Kandoth,1 Michael C. Wendl,1,2 William Schierding,1 Daniel C. Koboldt,1 Thomas B. Mooney,1 Matthew B. Callaway,1 David Dooling,1 Elaine R. Mardis,1,2,3 Richard K. Wilson,1,2,3 and Li Ding1,2,5

- Machine learning and classification (PolyPhen-2, CHASM, …)
  
  **Predicting the functional impact of protein mutations: application to cancer genomics**
  
  Boris Reva*, Yevgeniy Amzel1

  **Predicting Deleterious Amino Acid Substitutions**
  
  Pauline C. Ng1,2 and Steven Henikoff1,3,4

  **A method and server for predicting damaging missense mutations**
  
  Ivan A. Adzhubei1,7, Gerasimova5, Peer2

  **Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations**
  
  Hannah Carter1, Sining Chen2,3 Leyla Isik1, Svitlana Tyekucheva3, Victor E. Velculescu4, Kenneth W. Kinzler1, Bert Vogelstein4, and Rachel Karchin1