The spectra of somatic mutations across many tumor types

Mike Lawrence
Broad Institute of Harvard and MIT

1st Annual TCGA Scientific Symposium
November 17, 2011
mutation rates across cancer

Hematologic Childhood

Carcinogens

mutation rates across cancer

Hematologic Childhood

Carcinogens

mutation rates across cancer

Hematologic Childhood

Carcinogens

mutation rates across cancer

Hematologic Childhood

Carcinogens
mutation type

- C → T
- C → A
- C → G
- A → G
- A → T
- A → C

OV
mutation type
C → T
C → A
C → G
A → G
A → T
A → C

mutation context
T_T T_C T_A T_G
C_T C_C C_A C_G
A_T A_C A_A A_G
G_T G_C G_A G_G

transversions

transitions

right column
3’ base = G

mutation rate
(per million sites)

OV
GBM mutation type

C → T
C → A
C → G
A → G
A → T
A → C

mutation rate (per million sites)

left column 3' base = G

transversions

right column 3' base = G

transitions

GBM
LUSC lung squamous mutation type

- C → T
- C → A
- C → G
- A → G
- A → T
- A → C

Mutation context:
- T-T
- T-C
- T-A
- T-G
- C-T
- C-C
- C-A
- C-G
- A-T
- A-C
- A-A
- A-G
- G-T
- G-C
- G-A
- G-G

Transitions:
- C:G
- A:T

Transversions:
- LUSC lung squamous
<table>
<thead>
<tr>
<th>Mutation Context</th>
<th>Mutation Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_T</td>
<td>C → T</td>
</tr>
<tr>
<td>T_C</td>
<td>C → A</td>
</tr>
<tr>
<td>T_A</td>
<td>C → G</td>
</tr>
<tr>
<td>T_G</td>
<td>A → G</td>
</tr>
<tr>
<td>C_T</td>
<td>A → T</td>
</tr>
<tr>
<td>C_C</td>
<td>A → C</td>
</tr>
<tr>
<td>C_A</td>
<td>A → G</td>
</tr>
<tr>
<td>C_G</td>
<td>A → T</td>
</tr>
<tr>
<td>A_T</td>
<td>C → T</td>
</tr>
<tr>
<td>A_C</td>
<td>C → G</td>
</tr>
<tr>
<td>A_A</td>
<td>A → G</td>
</tr>
<tr>
<td>A_G</td>
<td>A → T</td>
</tr>
<tr>
<td>G_T</td>
<td>C → T</td>
</tr>
<tr>
<td>G_C</td>
<td>C → A</td>
</tr>
<tr>
<td>G_A</td>
<td>C → G</td>
</tr>
<tr>
<td>G_G</td>
<td>A → G</td>
</tr>
</tbody>
</table>

LUAD lung adeno

Bar graph showing the distribution of mutations by context and type.
Melanoma mutation type:

- C → T
- C → A
- C → G
- A → G
- A → T
- A → C

Mutation context:

- T-T
- T-C
- T-A
- T-G
- C-T
- C-C
- C-A
- C-G
- A-T
- A-C
- A-A
- A-G
- G-T
- G-C
- G-A
- G-G

Transitions:

- C:G

Transversions:

- A:T

Melanoma
mutation type

- $C \rightarrow T$
- $C \rightarrow A$
- $C \rightarrow G$
- $A \rightarrow G$
- $A \rightarrow T$
- $A \rightarrow C$

back row
5' base = T

transitions

bladder
finding significantly mutated genes
MutSig
scoring algorithm
MutSig
scoring algorithm

assume background mutation rate is:
- uniform across sequence contexts
- uniform across patients
- uniform across genes

version 0

patients

tally

significance
MutSig
scoring algorithm version 1

assume background mutation rate is:
· variable across sequence contexts
· uniform across patients
· uniform across genes
MutSig
scoring algorithm version 1

assume background mutation rate is:
- variable across sequence contexts
- uniform across patients
- uniform across genes

melanoma patients

gene W

gene X

gene Y

gene Z
MutSig
scoring algorithm _version 2_

assume background mutation rate is:
  · variable across sequence contexts
  · variable across patients
  · uniform across genes

patient 1  low mutation rate  patient 2  _high_ mutation rate
MutSig
scoring algorithm
version 2

assume background mutation rate is:
- variable across sequence contexts
- variable across patients
- uniform across genes

patients

tally

significance

patient 1
low mutation rate

gene A

gene B

gene C

gene D

patient 2
high mutation rate

*
Problem: mutation rate is heterogeneous across genes

Assume background mutation rate is:
- variable across sequence contexts
- variable across patients
- uniform across genes

MutSig
scoring algorithm version 2

patients → tally → significance

Problem: mutation rate is heterogeneous across genes

gene J

gene K

gene L

gene M
Problem: mutation rate is heterogeneous across genes

- average = 3 / Mb
- uniform across genes

MutSig scoring algorithm version 2

Assume background mutation rate is:
- variable across sequence contexts
- variable across patients
- uniform across genes
Problem: mutation rate is heterogeneous across genes

average = 3 / Mb

uniform across genes

MutSig scoring algorithm version 2

assume background mutation rate is:
· variable across sequence contexts
· variable across patients
· uniform across genes
Problem: mutation rate is heterogeneous across genes

average = 3 / Mb

uniform across genes

→ hits

Assume background mutation rate is:
- variable across sequence contexts
- variable across patients
- uniform across genes

MutSig scoring algorithm version 2

average = 3 / Mb

uniform across genes
Problem: mutation rate is heterogeneous across genes

25% genes: rate = 6 / Mb

average = 3 / Mb

uniform across genes

q<0.01

→ hits
Problem: mutation rate is heterogeneous across genes

Assume background mutation rate is:
- variable across sequence contexts
- variable across patients
- uniform across genes

\[
\begin{align*}
75\% \text{ genes: } & \text{rate} = 2 / \text{Mb} \\
25\% \text{ genes: } & \text{rate} = 6 / \text{Mb}
\end{align*}
\]
Problem: mutation rate is heterogeneous across genes

- 75% genes: rate = 2 / Mb
- 25% genes: rate = 6 / Mb

Assume background mutation rate is:
- variable across sequence contexts
- variable across patients
- uniform across genes

Average rate:
- uniform across genes = 3 / Mb
- 75% genes: rate = 2 / Mb
- 25% genes: rate = 6 / Mb

→ hits

q < 0.01

→ false positives
Lung cancer

457 patients
  180 lung squamous cell carcinoma
  277 lung adenocarcinoma
average mutation rate = 10 / Mb

MutSig results version 0
(assuming uniform background mutation rate across genes)

#1 * TP53
#2 * KRAS
#7  OR4A15
#13 * KEAP1
#14  OR8H2
#15 * STK11
#17  OR2T4
#25  OR2T3
#31  OR2T6
#48  CSMD3
#49  OR5D16
#55  RYR2
#100 CSMD1
#139 * PIK3CA
#158 RYR3
#159 MUC16
#161 OR2T33
#169 * NFE2L2
#172 OR10G8
#180 OR2L8
#198 MUC17
#217 TTN

* known lung cancer genes

all of these genes are extremely significant (q<10^-7)

total of 843 genes significantly mutated (q<0.01)

Bryan Hernandez
Peter Hammerman
Marcin Imielinski
Matthew Meyerson
Lung cancer

457 patients
- 180 lung squamous cell carcinoma
- 277 lung adenocarcinoma
Average mutation rate = 10 / Mb

MutSig results
(version 0)
(assuming uniform background mutation rate across genes)

- Total of 843 genes significantly mutated (q<0.01)
  - All of these genes are extremely significant (q<10^-7)

Known lung cancer genes

- TP53
- KRAS
- KEAP1

"Fishy" genes

- Olfactory receptors
  - 146 with q<0.01

- "Cub and sushi" proteins
  - Reported to be tumor suppressors
  - Significantly mutated in almost every tumor type
    (including TCGA ovarian)

- Ryanodine receptors
  - Cardiac calcium channels

- Mucins
  - Gel-forming proteins

- Titin
  - Largest human protein
  - 100x bigger than p53
  - 34,350 amino acids
  - 100 Kb coding sequence
Problem: mutation rate is actually heterogeneous across genes

Challenge: predict gene-specific background mutation rates

We eventually want to learn the background mutation rate of every gene (and all possible mutations at all basepairs!)

As we sequence more and more samples, we get closer to this goal.
Highly expressed genes have lower mutation rates

background mutation rate varies ten-fold or more across the genome

shown:
noncoding mutation rate from TCGA lung cancer dataset
Early-replicating genes have lower mutation rates

chr10

background mutation rate varies ten-fold or more across the genome

shown: noncoding mutation rate from TCGA lung cancer dataset

highly correlated

replication time also varies greatly across the genome

Sunyaev Lab (Harvard/BWH)
shown: replication time measurements from Chen et al. (2010) Genome Research 20:447
Late replication explains most olfactory receptors

16 ORs

chr1 (Mb)

mutations / Mb

late replication

early

All Genes

Early

Late

Olfactory Receptors
chr8

position (Mb)

mutations / Mb

replication

late

early

extrapolate even later replication times

CSMD3
initial model assumed a flat mutational landscape
landscape is actually *not* flat
improve estimate by binning together similar genes...
...or by local regression

- mutation rate
- average outward until neighborhood becomes too different from starting point
- gene A
- gene B
- gene expression level
- gene replication time
Lung cancer

MutSig v0
assuming uniform bkgd mutation rate across all genes

q<10^-7

843 genes significantly mutated (q<0.01)

* known lung cancer genes
"fishy" genes

#1 * TP53
#2 * KRAS
#7 OR4A15
#13 * KEAP1
#14 OR8H2
#15 * STK11
#17 OR2T4
#25 OR2T3
#31 OR2T6
#48 CSMD3
#49 OR5D16
#55 RYR2
#100 CSMD1
#139 * PIK3CA
#158 RYR3
#159 MUC16
#161 OR2T33
#169 * NFE2L2
#172 OR10G8
#180 OR2L8
#198 MUC17
#217 TTN
Lung cancer

MutSig v0 assuming uniform background mutation rate across all genes

843 genes significantly mutated (q<0.01)

q<10⁻⁷

improved MutSig using gene-specific background mutation rates

52 genes significantly mutated (q<0.01)

q<10⁻⁵

q~0.2

q=1

*known lung cancer genes

"fishy" genes

* most significant olfactory receptor

843 genes significantly mutated (q<0.01)

#1 * TP53
#2 * KRAS
#7 * KEAP1
#13 * OR4A15
#14 * OR8H2
#15 * STK11
#17 * OR2T4
#25 * OR2T3
#31 * OR2T6
#48 * CSMD3
#49 * OR5D16
#55 * RYR2
#100 * CSMD1
#139 * PIK3CA
#158 * OR8H2
#159 * MUC16
#161 * OR2T33
#169 * NFE2L2
#172 * OR10G8
#180 * OR2L8
#198 * MUC17
#217 * TTN

* STK11 #1
* NFE2L2 #4
* TP53 #7
* KRAS #8
* KEAP1 #11
* PIK3CA #12
* OR8H2 #181
* OR5T2 #276
* OR10J3 #334
* CSMD3 #388
* MUC16 #2614
* RYR2 #2898
* CSMD1 #4482
* OR2L8 #4825
* TTN #5650
* MUC17 #11496
## Correcting for variation in mutation rate

<table>
<thead>
<tr>
<th>Tumor Type</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Squamous</td>
<td>261 (50 OR)</td>
<td>18 (0 OR)</td>
</tr>
<tr>
<td>Lung Adeno</td>
<td>511 (93 OR)</td>
<td>33 (1 OR)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>177 (7 OR)</td>
<td>61 (0 OR)</td>
</tr>
<tr>
<td>Prostate</td>
<td>3 (0 OR)</td>
<td>3 (0 OR)</td>
</tr>
<tr>
<td>DLBCL</td>
<td>32 (1 OR)</td>
<td>15 (0 OR)</td>
</tr>
</tbody>
</table>

**Ultimate solution:** Learn the background rate
putting it all together
\[ \mu_o \cdot F_{p,s,c} = \mu_o \cdot F_p \cdot F_g \sum_k (w_{p,k} \cdot v_{k,c}) = \mu_{p,s,c} \]

- Overall mutation rate across entire dataset
- Relative mutation rate of gene \( g \) in patient \( p \) category \( c \)
- Relative mutation rate of patient \( p \)
- Sum across \( k \) "factors" (i.e. mutational processes)
- Weight of factor \( k \) in patient \( p \)
- Mutation rate in gene \( g \) patient \( p \) category \( c \)
significantly mutated genes across tumor types
Acknowledgements

MutSig Team
Petar Stojanov
Bryan Hernandez
Marcin Imielinski
Peter Hammerman
Gregory Kryukov
Eran Hodis
Chip Stewart

Analysis Team
Kristian Cibulskis
Andrey Sivachenko
Scott Carter
Gordon Saksena
Yotam Drier
Alex Ramos
Aaron McKenna
Rui Jing
Lihua Zou
David DeLuca
Elena Helman
Jaegil Kim
Cheng-Zhong Zhang
Sylvan Baca
Trevor Pugh
Nam Pho
Andrew Cherniak
Alex Kostic
Peter Chen
Nikhil Wagle

NHGRI   NCI

The Cancer Genome Atlas

Gaddy Getz
Stacey Gabriel
Levi Garraway
Lynda Chin
Matthew Meyerson
Eric Lander
Todd Golub

Broad Institute of Harvard and MIT

Dana Farber Cancer Institute

Shamil Sunyaev
Paz Polak
Leonid Mirny
Rameen Beroukhim
Steve Schumacher
Barbara Tabak
Cathy Wu
Jennifer Brown
Lili Wang
Youzhong Wan
Dan-Avi Landau
Aviv Regev
Nathalie Pochet
Project Management
Carrie Sougnez
Erica Shefler
Elizabeth Nickerson
Daniel Auclair
Marisa Cortes
Kristin Thompson

Adam Bass
Austin Dulak

Jim Robinson
Helga Thorvaldsdottir
Marc-Danie Nazaire
Jill Mesirov

Broad Institute Sequencing
Program and Platform
Tim Fennell
Sara Chauvin
Lauren Ambrogio
Sheila Fisher
Joshua Levin
Xian Adiconis
Andreas Gnirke
David Jaffe
Toby Bloom
Chad Nusbaum

Broad Institute Biological
Genetic Analysis Platform
Robb Onofrio
Brendan Blumenstiel
Huy Nguyen
Mellisa Parkin
Wendy Winckler

Broad Institute Biological
Samples Platform
Kristin Ardlie

Mike Berger
Mike Chapman
Jens Lohr
Derek Chiang
Craig Meruelo
Nicolas Stranks
Roel Verhaak
Barbara Weir
Shantanu Banerji

Firehose
Doug Voet
Mike Noble
Pei Lin
Dan DiCara
Lee Lichtenstein
Robert Zupko
Peter Carr

(cont.)

Mike Berger
Mike Chapman
Jens Lohr
Derek Chiang
Craig Meruelo
Nicolas Stranks
Roel Verhaak
Barbara Weir
Shantanu Banerji

Firehose
Doug Voet
Mike Noble
Pei Lin
Dan DiCara
Lee Lichtenstein
Robert Zupko
Peter Carr

(cont.)