The spectrum of mutations and mutated cancer genes across many tumor types

Julian Hess
Mike Lawrence
Gaddy Getz

Goals of cancer genome projects

(1) Novel cancer genes and pathways (oncogenes and tumor suppressors)
(2) A molecular classification scheme for cancer
(3) Potential therapeutic targets
(4) Understand the biology of cancer

Finding significant genes

Fundamental challenge: Distinguishing “driver” from “passenger” alterations

Model background mutational processes
Identify genes/regions/pathways with more mutations than predicted by the background model
candidate “driver” events or inaccurate background model systematic artifacts in mutation calling

MutSig: Approach -- Score genes according to number and types of mutations

Most mutations are in the "long tail" of genes

GBM (n=84)

Lung Adenocarcinoma (n=188)

Head & Neck (n=74)
Stransky, Egloff, Tward et al, Science (2011)

Ovarian (n=316)

Most mutations are in the "long tail" of genes

TP53

CDKN2A

BRAF, PIK3CA, KRAS, NRAS

+ BRAF, PIK3CA, KRAS, NRAS

Most mutations are in the "long tail" of genes

TP53

TP53

TP53

TP53

TP53

TP53

Most mutations are in the "long tail" of genes

TP53

TP53

TP53

TP53

TP53

TP53

TP53

Most mutations are in the "long tail" of genes

TP53

TP53

TP53

TP53

TP53

TP53

TP53
Mutational heterogeneity

Mutation rates across cancer (3078 WES, 27 tumor types) >1000-fold variation in mutation frequency among cancers

what factors contribute to variation of mutation rate across genome?

Modeling background mutation rates
Evidence of different mechanisms that affect mutation rates

DNA replication doesn't happen all at the same instant

Figure 4: A diagrammatic representation of replication timing in a 70-Mb segment of human chromosome 2. The red horizontal line represents time in S-phase, from early (top) to late (bottom). Grey data points each represent a different DNA sequence position along the length of chromosome 2 as indicated on the x-axis, with more positive values on the y-axis indicating earlier replication. A smoothed blue line is drawn through the data to visualize the domains of different replication timing.

 wakes mutation rate

chr10

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

Late replication explains most olfactory receptors

"fishy" genes have low expression and late replication time
concordance of the various metrics

concordance across tissue types

learning the landscape
initial model: assume flat mutational landscape

assuming a flat landscape didn't work

naive approach: bin the space uniformly

local regression: weight according to topography
Problem: As sample size and/or mutation rate increases → significant gene list increases and contains ‘fishy’ genes

Analyzing whole-exome data from 178 lung squamous cell carcinoma samples (~10/Mb) → 450 genes (q<0.1)

- 101/450 olfactory receptors
- 20% of 83 genes >4,000aa (p<10^{-11})
  e.g. TTN, MUC4/16/17, RYR2/3, ...
- 15% of 73 genes >1Mb (p<10^{-5})
  e.g. CSMD1/3, NLRN51/4, PARK2

<table>
<thead>
<tr>
<th>Ovarian (n=316)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA network, Nature (2011)</td>
</tr>
</tbody>
</table>

Finding cancer genes across ~5000 tumor/normal pairs from 21 tumor types

1. Can we detect all currently known cancer genes?
2. Will we reveal new cancer genes?
3. Have we completed the catalog of all cancer genes (mutated >2% of patients)?

Crucial to correct for variation in background rates

450 genes (q<0.1) → 11 genes (q<0.1)

MutSigCV

<table>
<thead>
<tr>
<th>450 genes (q&lt;0.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 genes (q&lt;0.1)</td>
</tr>
</tbody>
</table>

| TP53 |
| KEAP1 |
| NFE2L2 |
| PIK3CA |
| PTEN |
| RB1 |
| MLH1 |
| NOTCH1 |
| FBXW7 |
| MHA |

~5000 tumors across 21 tumor types

<table>
<thead>
<tr>
<th>Tumor type</th>
<th>Tumor type code</th>
<th>No. of tumor-normal pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute myeloid leukemia</td>
<td>LAML</td>
<td>796</td>
</tr>
<tr>
<td>Bladder</td>
<td>UCEA</td>
<td>86</td>
</tr>
<tr>
<td>Breast</td>
<td>ESR1</td>
<td>56</td>
</tr>
<tr>
<td>Colorectal</td>
<td>CRC</td>
<td>293</td>
</tr>
<tr>
<td>Diffuse large B-cell lymphoma</td>
<td>DLBCL</td>
<td>98</td>
</tr>
<tr>
<td>Endometrial</td>
<td>UCEC</td>
<td>248</td>
</tr>
<tr>
<td>Esophageal adenocarcinoma</td>
<td>ESO</td>
<td>141</td>
</tr>
<tr>
<td>Glioblastoma multiforme</td>
<td>GBM</td>
<td>206</td>
</tr>
<tr>
<td>Head and neck</td>
<td>HGSC</td>
<td>184</td>
</tr>
<tr>
<td>Kidney clear cell</td>
<td>KIRC</td>
<td>417</td>
</tr>
<tr>
<td>Lung adenocarcinoma</td>
<td>LUAD</td>
<td>406</td>
</tr>
<tr>
<td>Lung squamous cell carcinoma</td>
<td>LSQ</td>
<td>178</td>
</tr>
<tr>
<td>Mesothelioma</td>
<td>MES</td>
<td>92</td>
</tr>
<tr>
<td>Melanoma</td>
<td>MEA</td>
<td>119</td>
</tr>
<tr>
<td>Multiple myeloma</td>
<td>MM</td>
<td>207</td>
</tr>
<tr>
<td>Neuroblastoma</td>
<td>NB</td>
<td>81</td>
</tr>
<tr>
<td>Ovarian</td>
<td>OV</td>
<td>339</td>
</tr>
<tr>
<td>Prostate</td>
<td>PRAD</td>
<td>138</td>
</tr>
<tr>
<td>Rhabdoid tumor</td>
<td>Rhabd</td>
<td>35</td>
</tr>
</tbody>
</table>

How to find cancer genes?

- Somatic mutations
  - Substitutions: ~3 million
  - Indels: ~80,000
  - Mutations/patient: 674
  - 93% of genes had at least 5 mutations in the dataset
Combined Analysis of ~5000 tumors

Three types of scores (MutSig suite)

- Nonsyn/syn
- Clustering
- Conservation

For each analysis:
- Use False Discovery Rate (FDR) < 0.1
- Check q-q plot lies on diagonal

Analysis of 21 tumor types →
336 gene x tumor type pairs (q<0.1) involving 227 cancer genes

ubiquitous cancer genes

NRAS
neuroblastoma RAS viral (v-ras) oncogene homolog

This is an N-ras oncogene encoding a membrane protein that shuttles between the Golgi apparatus and the plasma membrane. This shuttling is regulated through palmitoylation and depalmitoylation by the ZDHHC9-GOLGA7 complex. The encoded protein, which has intrinsic GTPase activity, is activated by a guanine nucleotide-exchange factor and inactivated by a GTPase activating protein. Mutations in this gene have been associated with somatic rectal cancer, follicular thyroid cancer, autoimmune lymphoproliferative syndrome, Noonan syndrome, and juvenile myelomonocytic leukemia.
PIK3CA  phosphoinositide-3-kinase, catalytic, alpha polypeptide

Phosphoinositide-3-kinase (PI3K) is composed of an 85 kDa regulatory subunit and a 110 kDa catalytic subunit. The protein encoded by this gene represents the catalytic subunit, which uses ATP to phosphorylate PtdIns, PtdIns4P and PtdIns(4,5)P2. This gene has been found to be oncogenic and has been implicated in cervical cancers.

RB1  retinoblastoma 1

The protein encoded by this gene is a negative regulator of the cell cycle and was the first tumor suppressor gene found. The active, hypophosphorylated form of the protein binds transcription factor E2F1. Defects in this gene are a cause of childhood cancer retinoblastoma (RB), bladder cancer, and osteogenic sarcoma.

CASP8  caspase 8, apoptosis-related cysteine peptidase

This gene encodes a member of the cysteine-aspartic acid protease (caspase) family. Sequential activation of caspases plays a central role in the execution-phase of cell apoptosis. Caspases exist as inactive proenzymes composed of a prodomain, a large protease subunit, and a small protease subunit. Activation of caspases requires proteolytic processing at conserved internal aspartic residues to generate a heterodimeric enzyme consisting of the large and small subunits. This protein is involved in the programmed cell death induced by Fas and various apoptotic stimuli.

APC  adenomatous polyposis coli

This gene encodes a tumor suppressor protein that acts as an antagonist of the Wnt signaling pathway. It is also involved in other processes including cell migration and adhesion, transcriptional activation, and apoptosis. Defects in this gene cause familial adenomatous polyposis (FAP), an autosomal dominant pre-malignant disease that usually progresses to malignancy. Disease-associated mutations tend to be clustered in a small region designated the mutation cluster region (MCR) and result in a truncated protein product.
**GATA3**

GATA binding protein 3

This gene encodes a protein which belongs to the GATA family of transcription factors. The protein contains two GATA-type zinc fingers and is an important regulator of T-cell development and plays an important role in endothelial cell biology. Defects in this gene are the cause of hypoparathyroidism with sensorineural deafness and renal dysplasia.

**VHL**

von Hippel-Lindau tumor suppressor

Von Hippel-Lindau syndrome (VHL) is a dominantly inherited familial cancer syndrome predisposing to a variety of malignant and benign tumors. A germline mutation of this gene is the basis of familial inheritance of VHL syndrome. The protein encoded by this gene is a component of the protein complex that includes elongin B, elongin C, and cullin-2, and possesses ubiquitin ligase E3 activity. This protein is involved in the ubiquitination and degradation of hypoxia-inducible-factor (HIF), which is a transcription factor that plays a central role in the regulation of gene expression by oxygen. RNA polymerase II subunit POLR2G/PBF1 is also reported to be a target of this protein.

**NPM1**

nucleophosmin (nucleolar phosphoprotein B23, numatrin)

This gene encodes a phosphoprotein which moves between the nucleus and the cytoplasm. The gene product is thought to be involved in several processes including regulation of the ARF/p53 pathway. A number of genes are fusion partners in acute myeloid leukemia. Mutations in this gene are associated with acute myeloid leukemia. More than a dozen pseudogenes of this gene have been identified. Alternative splicing results in multiple transcript variants.

**CDKN1A**

cyclin-dependent kinase inhibitor 1A

This gene encodes a potent cyclin-dependent kinase inhibitor. The encoded protein binds to and inhibits the activity of cyclin-CDK2 or -CDK4 complexes, and thus functions as a regulator of cell cycle progression at G1. The expression of this gene is tightly controlled by the tumor suppressor protein p53, through which this protein mediates the p53-dependent cell cycle G1 phase arrest in response to a variety of stress stimuli. This protein can interact with proliferating cell nuclear antigen (PCNA), a DNA polymerase accessory factor, and plays a regulatory role in S phase DNA replication and DNA damage repair. This protein was reported to be specifically cleaved by CASP3-like caspases, which thus leads to a dramatic activation of CDK2, and may be instrumental in the execution of apoptosis by following caspase activation.
**ATRX**

alpha thalassemia/mental retardation syndrome X-linked

The protein encoded by this gene contains an ATPase/helicase domain, and thus it belongs to the SWI/SNF family of chromatin remodeling proteins. The mutations of this gene are associated with an X-linked mental retardation (XLMR) syndrome most often accompanied by alpha-thalassemia (ATRX) syndrome. These mutations have been shown to cause diverse changes in the pattern of DNA methylation, which may provide a link between chromatin remodeling, DNA methylation, and gene expression in developmental processes. The protein is found to undergo cell cycle-dependent phosphorylation, which regulates its nuclear matrix and chromatin association, and suggests its involvement in the gene regulation at interphase and chromosomal segregation in mitosis.

**ELF3**

E74-like factor 3 (ets domain transcription factor, epithelial-specific)

Transcriptional activation that binds and transactivates ETS sequences containing the consensus nucleotide core sequence GGA[AT]. Acts synergistically with POU2F3 to transactivate the SPRR2A promoter and with RUNX1 to transactivate the ANGPT1 promoter. Also transactivates collagenase, CCL20, CLND7, KRT8, NOS2, PTGS2, SPRR2B, TGFBR2 and TGM3 promoters. Represses KRT14 promoter activity. Involved in mediating vascular inflammation. May play an important role in epithelial cell differentiation and tumorigenesis. May be a critical downstream effector of the ERBB2 signaling pathway.

**Novel candidate cancer genes**

- 4 anti-proliferative (LoF)
- 6 proliferation (GoF, recurrent mutations)
- 5 pro-apoptotic (LoF)
- 6 genome stability
- 5 chromatin regulation
- 3 immune evasion (LoF, HLA-B)
- 3 RNA processing
- 1 protein homeostasis (E3 ligase, recurrent mutations)

33 novel genes with compelling evidence

**Example: RHEB**

Small GTPase (Ras-homolog enriched in brain)

The TSC1-TSC2 complex is regulated by many oncogenes and tumor suppressors, including mTORC1 and mTORC2. This complex inhibits the activity of RHEB, a small GTPase that activates the mTORC1 complex. Y35N is a critical amino acid in the effector domain of RHEB.
**Example: RHOA**

Small GTPase (part of Ras superfamily)

5 instances of E40Q
1 instance of Y42I
in effector domain

---

**Example: RAD21**

Involved in repairing double strand breaks and chromatid cohesion

Significantly mutated in AML
RAD21 partners: SMC1A, SMC3
previously known to be significant in AML

---

**Example: PCBP1**

Blocks translation of certain genes by binding to polyC regions

Hotspot at two leucines (Leu100, Leu102)
in region mediating dimerization of KH domains

---

**Downsampling analysis**

Assessing the pace of discovery
Down-sampling analysis

What if we had only 100 patients?

No longer found:

*DTAC1
*24S
*125L
*1210A
*M10
*M2451
Down-sampling analysis

What if we had only 50 patients?

No longer found:

- MDR1
- CYP3A4
- ABCB1
- EGFR
- EGFR
- STAG2
- NGS1
- SET2
- NRAS
- PENK
- FBXW7
- MIR15A
- MIR16
- TSC22
- PDCD5
Down-sampling analysis

Saturation analysis: Down-sampling within tumor types shows steep rise

Saturation analysis: Down-sampling across tumor types shows steep rise

Saturation analysis by frequency class

No longer found:

≥ 20% largely discovered
<20% still rising rapidly
How many more patients do we need to analyze?

PanCan14k (ongoing!)
4700 to 14000 patients
Summary

1. Cancer mutation datasets are complex, with heterogeneity at all levels of the data.

2. In a typical tumor type, there are:
   a few genes mutated at high frequency
   many genes mutated at lower frequencies

3. High-frequency genes:
   account for only a small fraction of all driver mutations
   have nearly all been discovered

4. Lower-frequency genes:
   account for the vast majority of all driver mutations
   affect nearly all patients
   are still being discovered at a rapid pace
   are crucial to understand in the fight against cancer
Acknowledgements

Analysis Team

Petar Stojanov
Paz Polak
Kristian Cibulskis
Scott Carter
Andrey Sinitsenko
Adam Kiezun
Craig Kumer
Yotam Drier
Gordon Sakkena
Miran Inselenski
Lihua Zou
Alex Ramos
Mike Berger
Trevor Pugh
David DeLuca
Mike Chapman
Aaron McKenna
Gregory Krylov
Alex Kostic
Nicolas Stroinsky
Derek Chiang
Ray Verhaak
Shantaxi Benerji
Chip Stewart
Nikhil Wagle

Firehose
Doug Voet
Pei Lin
Lee Lichtenstein
Rakim Singh
Peter Carr

GGAC
Mike Noble
Robert Zupke
Dan DiCarra
Ha Li Zhang
Juck Cho
Spring Liu

Rameen Bernsuhkim
Amara Taylor-Wanu
Svena Schumacher
Barbara Babak
Adam Bass
Austin Dukak

Project Management
Carrie Singney
Eric Sheffer
Elizabeth Nickerson
Daniel Auclair
Maria Cortes
Kristin Thompson

Jim Robinson
Helga Thorkelsdottir
Mark Danis-Neauser
Jill Mesirov

Broad Institute
Genomics Platform
Stacey Gabriel
Tim Fennell
David Jaffe
Toby Bloom
Chad Nusbaum
Wendy Winkler
Kristin Ardlie
Jane Wilkinson
Sara Chaurin
Lauren Ambrogio
Sheila Fisher
Joshua Levin
Xian Adinonis
Andreas Gavrke
Robb Onofrio
Brendan Blumenstiel
Huy Nguyen
Melissa Parkin

Mark DePristo
Eric Banks
Kiran Garmella