COMPUTING the CURE of CANCER

Jiguang Wang

Division of Life Science
Department of Chemical and Biological Engineering

The Hong Kong University of Science and Technology
In Hong Kong

Statistics by Department of Health, Hong Kong
cancer

a malignant growth or tumor, uncontrolled cell division, growth or tumor.
Chronic Myeloid Leukemia
(a type of blood cancer)
Targeted therapy—Step ONE: Identify the cancer CELLS
Targeted therapy—Step TWO: Identify a **PROTEIN**
Targeted therapy—Step THREE: Identify a CURE
Targeted Therapy in Chronic Myeloid Leukemia

- Identify targets
- Identify Drug
- Deliver Drug

Sample Sequencing

More precise Treatment

Literature Prediction

Bcl-Abl

Age-Standardised Ten-Year Net Survival, Selected Cancers, Adults (Aged 15-99), England and Wales, 2010-2011

Ten-year survival for 2005-2006 and 2010-2011 is predicted using an excess hazard statistical model.

Survival for bowel cancer is a weighted average derived from data for colon (C18) and rectum cancer (C19 C20, C21.8).

Source: cruk.org/cancerstats

You are welcome to reuse this Cancer Research UK statistics content for your own work. Credit us as authors by referencing Cancer Research UK as the primary source. Suggested style: Cancer Research UK, full URL of the page, Accessed [month] [year].
All fast-growing cells
Chemotherapy

Cells with particular mutations
Targeted therapy

Nature Medicine, Clevers 2011
KEY Questions

How to identify a target?

How to identify a drug?
To identify a target?
Cost of sequencing
Next Generation Sequencing: Main commercialized technologies

- **Illumina**
  - MiSeq
  - HiSeq 500
  - HiSeq 2500
  - HiSeq X Ten

- **PacBio RS**

- **454** – pyrosequencing
“The way we do RNA-seq now is... you take the transcriptome, you blow it up into pieces and then you try to figure out how they all go back together again... If you think about it, it’s kind of a crazy way to do things.”

Michael Snyder
Stanford University


Mutations in cancer

• Types of genome alterations that can be detected by next-generation sequencing:

Meyerson et al, Nature Rev Genetics 2010
Single Nucleotide Variants
Single Nucleotide Variants in general population

ATTGCAATTCGTGG...ATCGAGCCA...TACGATTGCAACGCCG...
ATTGCAAGCCGTGG...ATCTAGCCA...TACGATTGCAAGCCG...
ATTGCAAGCCGTGG...ATCTAGCCA...TACGATTGCAAGCCG...
ATTGCAATTCGTGG...ATCGAGCCA...TACGATTGCAACGCCG...
ATTGCAAGCCGTGG...ATCTAGCCA...TACGATTGCAAGCCG...
ATTGCAATTCGTGG...ATCGAGCCA...TACGATTGCAACGCCG...
Single Nucleotide Variants in Cancer Cells

TP53
C>T  R248Q
INDELs in Cancer Cells

Normal Cell

Tumor Cell

Chr17

TP53

TC/- Frame shift alteration
TOBI: Tumor-Only Boosting Identification of Germline and Somatic Cancer Drivers

WES of a Cohort of Patients

Training set:
Tumors with known somatic variants (5-20 cases)

Testing set:
Tumors without matched normals (any number of cases)

Variant Calling & Annotation

O(1,000,000) Variants

Technical & Biological Filtering

O(100) Variants

Supervised ML

O(10) Predicted Somatic

Machine Learning

O(10) Predicted Somatic

npj Genomic Medicine, in press, 2017
sequencing error rates (parameter for binomial distribution)

\[ X \sim \text{Bin}(n, p) \]

\[ \pi(p|\alpha, \beta) = \text{Beta}(\alpha, \beta) = \frac{p^{\alpha-1}(1 - p)^{\beta-1}}{\text{B}(\alpha, \beta)} \]
normal reference DNAs
(collected from other patients than the target)

sequencing error rate predictive distribution for each position

target normal DNAs
(for filtration of germline mutations)

target tumor DNAs

the observed mismatch ratio is significantly deviated from the predictive sequencing error distribution (cannot be explained by sequencing errors)

true somatic mutation!
## One Example of Variant Calling

- A patient was diagnosed with GBM
- Surgery was performed
- Both cancer cells and normal blood were sent for Whole Exome Sequencing

| Chromosome | Position | Ref | Alt | ID (SNP) | ID (COSM) | Allele Histotype | Codon Change | Amino Acid Change | Gene Name | Codon pm50 | Codon pm100 | Codon pm1000 | Codon pm10000 | Exon | Total Variants | Somatic Variants |
|------------|----------|-----|-----|----------|-----------|-----------------|--------------|------------------|-----------|-------------|------------|-------------|--------------|---------------|----------|----------------|-----------------|
| chr4       | 8594643  | C   | G   | rs768922132 | COSM1422270 | NONSENSE       | c/C>cGc      | R661*           | EOM5      | 157         | 68          | 1             | 0             | ~93,007       | 64         | significant   | somatic         |
| chr6       | 27758369 | A   | C   | rs7770371  | COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr7       | 11116810 | A   | C   | rs77037019 | COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr8       | 38435221 | A   | C   | rs773708276| COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr10      | 12263428 | G   | A   | rs74704136 | COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr11      | 38435221 | A   | C   | rs773708276| COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr12      | 11116810 | A   | C   | rs77037019 | COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr14      | 2436497  | T   | C   | rs200095671| COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr15      | 18352088 | G   | T   | rs37141656 | COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr16      | 55518390 | T   | C   | rs37141656 | COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr17      | 18317704 | G   | T   | rs772790612| COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr18      | 21422680 | C   | T   | rs772790612| COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr19      | 64392555 | T   | C   | rs755899214| COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
| chr20      | 56935875 | T   | C   | rs37141656 | COSM1646823 | NONSENSE       | c/G>cGc      | G331H          | TP53      | 143         | 107         | 3            | 3             | ~93,007       | 64         | significant   | somatic         |
Mutations in cancer

• Types of genome alterations that can be detected by next-generation sequencing:

Meyerson et al, Nature Rev Genetics 2010
Copy number variants

- Deletion: One copy of "C"
- Duplication: Three copies of "C"
Pipeline of Copy number data analysis based on whole-exome sequencing data

BAM test files
BAM \( T_1 \), BAM \( T_2 \), ...

BAM control files
BAM \( C_1 \), BAM \( C_2 \), ...

RC test data
RC \( T_1 \), RC \( T_2 \), ...

RC control data
RC \( C_1 \), RC \( C_2 \), ...

NRC test data
NRC \( T_1 \), NRC \( T_2 \), ...

NRC control data
NRC \( C_1 \), NRC \( C_2 \), ...

BAM file filtering and RC calculations

Normalization

Pooling

Somatic

HSLM and FastCall

Results
BED, VCF

Genome Biology 2013
Copy number Change of GBM
Mutations in cancer

- Types of genome alterations that can be detected by next-generation sequencing:

Meyerson et al, Nature Rev Genetics 2010
Chimeric Transcript Detection

Abate et al., *Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model*. Bioinformatics, 2012
Mutations in cancer

- Types of genome alterations that can be detected by next-generation sequencing:

Meyerson et al, Nature Rev Genetics 2010
Virus detection in cancer

Sequencing Reads

HBV (Hepatitis B Virus)

Unmatched Reads
Application in brain tumors
Glioblastoma (GBM)

• GBM is the most **Common** and most **Aggressive** primary brain tumor in adults

• The incidence of GBM is 2~3 per 100,000 people per year

• Without treatment, survival is only **3 months**

• Even under the most intensive therapy, most patients will die in **12~15 months**
Application I: Treatment Naïve brain tumors

Primary GBM

Secondary GBM

Recurrent GBM

Low-grade glioma

Treatment naïve brain tumors
Genomic Landscape of GBM at Diagnosis

- Cross-sectional studies (~200 patients) have identified many driver genes:
  
  **Targetable Fusion FGFR3-TACC3**

  *Science 2012;337:1231-1235.*
Genomic Landscape of GBM

Brennan et al. Cell 2013
Pan-glioma study

1122 glioma samples

Mutations
DNA Methylation
CNV
Gene Expression
Protein Expression

IDHmutant
G-CIMP-high
G-CIMP-low
Codel

Epigenetic Signatures

IDHwild-type
Classic-like
LGm6-GBM
Mesenchymal-like
PA-like LGG

Ceccarelli et al. Cell 2016
Application II: Recurrent Glioblastoma (GBM)

Primary GBM -> More Years -> Recurrent GBM

Low-grade glioma -> More Years -> Secondary GBM
Recurrent GBM with Matched Primary Tumor

Somatic Mutations

Istituto Neurologico Carlo Besta
MD Anderson
TCGA
UCSF
Kyoto University
Samsung Medical Center

Initial
Common
Recurrence
Unknown

Recurrent GBM with Matched Primary Tumor

Somatic Mutations

Primary GBM
JM2
MMR
Hypomutated

TP53
ARID1B
EGFR
PTEN
PIK3CA
PIK3R1
PIK3CG
PDGFRA
RB1
APC
PTEN

MLH1
CDK4
EGFRvIII
MMR
MSI
PDGFRβ
PTP1B
ATRX

DSL
Mutation Landscape of Recurrent GBM

EGFRvIII: Deletion of *EGFR* exon 2-7

Novel MGMT fusions in Recurrent GBM

Wang, Nature Genetics 2016
Recurrent GBM in a different location
**Spatiotemporal Landscape of 52 Glioblastoma Patients (127 samples)**

<table>
<thead>
<tr>
<th>TimePoint</th>
<th>Initial</th>
<th>Recur</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion</td>
<td>ATRX</td>
<td>EGFR</td>
</tr>
<tr>
<td></td>
<td>FGFR3-TACC3</td>
<td>MGMT</td>
</tr>
<tr>
<td>SNV</td>
<td>Clonal</td>
<td>SubClonal</td>
</tr>
<tr>
<td>CNV</td>
<td>Amplification</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

**SubClonal**

**Neutral**

**Deletion**

**Long-Time Recurrence**

---

A Multiverse Model of GBM Evolution

1. Uniformly high ITH and subclone mixing in even distant regions

2. Most detectable ITH originates from early private alterations not from later clonal expansions


Application III: secondary Glioblastoma (GBM)

- Primary GBM
- Low-grade glioma
- Recurrent GBM
- Secondary GBM

Time:

- Months
Genomic Landscape of Secondary GBM

~200 sGBM

Primary GBM

Secondary GBM

Low Grade Glioma

MET alteration

Unpublished
To identify a drug?
Drug Screen of Patient Derived Cell Line

Genomic Profiling

- Cancer type (14)
- Patients (358)
- Tissue (393)
- Whole exome seq (95)
- Whole exome seq blood (66)
- GliomaScan (41)
- GliomaScan blood (28)
- CancerScan (281)
- RNASeq (109)
- PDC (101)
- GilomaScan (80)
- RNASeq (50)

Gene-drug association

Clinical response prediction

Chemical response (60 drugs)

ATP-based cell survival analysis

Compute DRC

Calculate AUC

Spearman correlation

mRNA expression on Tissue

mRNA expression on PDC

Lee*, Liu*, Wang*, et al, Nature Genetics, under revision
Finding a drug to treat a recurrent GBM

An mTOR inhibitor was applied in GBM15

Finding a drug to treat secondary GBM

Unpublished
Targeting Founding Mutations

Lee*, Wang*, et al, Nature Genetics, 2017
The Truncal Target Hypothesis

Lee*, Wang*, et al, Nature Genetics 2017
Machine learning approach to find the pattern of gene-drug interaction

Gene-Drug Map from Literature
- Check correlation on GBM

Gene-Drug Map of GBM
- For each drug-target pair

Network diffusion kernel
- Get top 500 genes have high network connectivity with the direct drug target
- Overlapped by the genes have consist expression between PDC and Tissue

Fitting the elastic-net model
\[
\min_{(\beta_0, \beta) \in \mathbb{R}^{n+1}} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \left[ (1 - \alpha)\|\beta\|_2^2 / 2 + \alpha \|\beta\|_1 \right]
\]
- $K$ fold cross validation to optimize Beta, alpha, lambda
- Bootstrap 100 datasets (80% samples) to get robust estimation

Predictive genes additional to the drug target
Personalized Cancer Therapy

From MD Anderson website
Acknowledgements

Raul Rabadan Lab
Antonio Iavarone Lab

Emanuela Cazzato
Veronique Frattini
Daniel Rosenbloom
Sakellarios Zairis
Erik Ladewig

Do-Hyun Nam Lab

Jin-Ku Lee
Jason Sa
In-Hee Lee

Tao Jiang Lab

Zheng Zhao
Zheng Wang
Xing Liu
Huimin Hu

Matter (pan-cancer, AI technologies)

Wang Genomic Lab

Biaobin Jiang
Yiyun Chen
Quanhua Mu
Yoonhee Nam

Center of Systems Biology & Human Health

Life Science

BME

港科技大學

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY