Participating Experts:

Michael Dean, Ph.D.
Center for Cancer Research, NIH
Bethesda, MD

Stephan Züchner, M.D.
Hussman Institute for Human Genomics
Miami, FL

Pauline C. Ng, Ph.D.
J Craig Venter Institute
San Diego, CA
EXOME SEQUENCING AND CANCER

Dr. Michael Dean
National Cancer Institute
Frederick, MD
CANCER OVERVIEW

Normal → Pre-malignant → Malignant → Metastatic
CANCER OVERVIEW

Viruses
Pathogens
Chemicals
Radiation
Inflammation
Hormonal Stimulation

Normal
Pre-malignant
Malignant
Host variation
Somatic Mutation
Epigenetic Changes
Chromosome Rearrangement
Metastatic
CANCER ALTERATIONS

• Germline Mutations
  – High Penetrance (RB1, P53, BRCA1, PTCH, VHL…)
  – Low Odds Ratio variants/Modifiers
    • FGFR2 breast cancer
    • MSMB prostate cancer
    • 8q24 Breast, Prostate, Gastric, …
# CANCER ALTERATIONS

## Somatic Alterations
- Mutations
- Deletions
- Translocations
- Chromosome loss
  - uniparental disomy
- Gene Amplification
- Methylation

## RNA
- Expression
- Alternative splicing
- Trans-splicing
- Hybrid Transcripts
- RNA Editing
## NCI-60 CELL LINES

<table>
<thead>
<tr>
<th>Tissue Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>6</td>
</tr>
<tr>
<td>Lung</td>
<td>9</td>
</tr>
<tr>
<td>Colon</td>
<td>7</td>
</tr>
<tr>
<td>Brain</td>
<td>6</td>
</tr>
<tr>
<td>Melanoma</td>
<td>9</td>
</tr>
<tr>
<td>Ovarian</td>
<td>7</td>
</tr>
<tr>
<td>Kidney</td>
<td>8</td>
</tr>
<tr>
<td>Prostate</td>
<td>2</td>
</tr>
<tr>
<td>Breast</td>
<td>7</td>
</tr>
</tbody>
</table>

Shoemaker Nature Reviews Cancer 6, 813-823 2006
<table>
<thead>
<tr>
<th>Tumor Type</th>
<th>Count</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>6</td>
<td>Drug Response to &gt; 100,000 compounds</td>
</tr>
<tr>
<td>Lung</td>
<td>9</td>
<td>RNA Expression-multiple platforms</td>
</tr>
<tr>
<td>Colon</td>
<td>7</td>
<td>Functional/protein assays</td>
</tr>
<tr>
<td>Brain</td>
<td>6</td>
<td>CGH</td>
</tr>
<tr>
<td>Melanoma</td>
<td>9</td>
<td>Illumina 1M SNPs</td>
</tr>
<tr>
<td>Ovarian</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Kidney</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Shoemaker Nature Reviews Cancer 6, 813-823 2006
QUALITY CONTROL

Exome Variants vs. Illumina 1M SNPs
Compare predicted variants from exome to exonic SNPs obtained by genotyping.

<table>
<thead>
<tr>
<th>Tumor</th>
<th>Discrepancies</th>
<th>SNP Genotypes</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>786-0</td>
<td>18</td>
<td>2876</td>
<td>99.4%</td>
</tr>
<tr>
<td>SN12c</td>
<td>32/2692</td>
<td></td>
<td>98.8%</td>
</tr>
<tr>
<td>PC3</td>
<td>21/1610</td>
<td></td>
<td>98.7%</td>
</tr>
<tr>
<td>MCF7</td>
<td>11/1065</td>
<td></td>
<td>99%</td>
</tr>
</tbody>
</table>
QUALITY CONTROL

Exome Variants vs. Transcriptome variants
Compare predicted variants from exome to predicted variants in transcriptome.

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Gene</th>
<th>VarFreq</th>
<th>Depth</th>
<th>VarFreq</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>786-0</td>
<td>ATM</td>
<td>75%</td>
<td>4</td>
<td>100%</td>
<td>1</td>
</tr>
<tr>
<td>786-0</td>
<td>BRCA1</td>
<td>100%</td>
<td>7</td>
<td>100%</td>
<td>6</td>
</tr>
<tr>
<td>786-0</td>
<td>BRIP1</td>
<td>100%</td>
<td>7</td>
<td>100%</td>
<td>5</td>
</tr>
<tr>
<td>MCF7</td>
<td>BRIP1</td>
<td>92%</td>
<td>24</td>
<td>100%</td>
<td>18</td>
</tr>
<tr>
<td>PC-3</td>
<td>MSH2</td>
<td>50%</td>
<td>8</td>
<td>29%</td>
<td>14</td>
</tr>
<tr>
<td>PC-3</td>
<td>MSH2</td>
<td>50%</td>
<td>8</td>
<td>29%</td>
<td>14</td>
</tr>
<tr>
<td>SN12C</td>
<td>MSH2</td>
<td>40%</td>
<td>10</td>
<td>60%</td>
<td>10</td>
</tr>
<tr>
<td>Cell Line</td>
<td>Type</td>
<td>Variants</td>
<td>New</td>
<td>NS</td>
<td>Null</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>----------</td>
<td>------</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>786-0</td>
<td>Renal</td>
<td>3469</td>
<td>235</td>
<td>161</td>
<td>12</td>
</tr>
<tr>
<td>SN12c</td>
<td>Renal</td>
<td>3505</td>
<td>179</td>
<td>119</td>
<td>13</td>
</tr>
<tr>
<td>PC3</td>
<td>Prostate</td>
<td>2960</td>
<td>85</td>
<td>53</td>
<td>2</td>
</tr>
<tr>
<td>MCF7</td>
<td>Breast</td>
<td>2570</td>
<td>95</td>
<td>57</td>
<td>4</td>
</tr>
</tbody>
</table>

NS= Non-synonymous
OTHER APPLICATIONS

• Hybrid Transcripts
  – Predicted the presence of hybrid transcripts representing potential translocations, deletions, trans-splicing

• Expression
DIGITAL CGH/LOH ANALYSIS

Uniparental Disomy

Allelic Imbalance
786-0 SKY

786-0 Renal Ca SKY and Digital LOH Analysis
Galileo's letter to the Prince of Venice
<table>
<thead>
<tr>
<th>NCI-Frederick</th>
<th>454 Life Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike Nickerson</td>
<td>Zdenek Markovic</td>
</tr>
<tr>
<td>Kate McGee</td>
<td>Antonio Goncalves</td>
</tr>
<tr>
<td>Bert Gold</td>
<td>Qin Zhao</td>
</tr>
<tr>
<td>SAIC</td>
<td>Ben Boese</td>
</tr>
<tr>
<td>Meredith Yeager</td>
<td>Pascal Bouffard</td>
</tr>
<tr>
<td>Claudia Stewart</td>
<td>Lewyn Li</td>
</tr>
<tr>
<td>Roche Applied Science</td>
<td>Tim Harkins</td>
</tr>
</tbody>
</table>
Exome sequencing of a multigenerational human pedigree

Stephan Züchner, MD
Associate Professor of Human Genetics and Neurology
Genetic Variation in an Individual Human Exome

Pauline C. Ng*, Samuel Levy, Jiaqi Huang, Timothy B. Stockwell, Brian P. Walenz, Kelvin Li, Nelson Axelrod, Dana A. Busam, Robert L. Strausberg, J. Craig Venter

J. Craig Venter Institute, Rockville, Maryland, United States of America


Targeted capture and massively parallel sequencing of 12 human exomes

Sarah B. Ng¹, Emily H. Turner¹, Peggy D. Robertson¹, Steven D. Flygare¹, Abigail W. Bigham², Choli Lee¹, Tristan Shaffer¹, Michelle Wong¹, Arindam Bhattacharjee³, Evan E. Eichler¹,³, Michael Bamshad², Deborah A. Nickerson¹ & Jay Shendure¹
Eight exomes from a three-generational pedigree were studied

- Whole Human Exome Capture Arrays (Roche)
- 454 GS FLX sequencing platform
- GSMapper software package
Coverage of targeted exons at different amounts of aligned sequence

~1Gb of aligned sequence:
- ≥3x coverage: 94% of targets
- ≥10x coverage: 56% of targets
- ≥20x coverage: 13% of targets

~2Gb of aligned sequence:
- ≥3x coverage: 98% of targets
- ≥10x coverage: 86% of targets
- ≥20x coverage: >50% of targets
Variant detection in eight exomes

Based on the conservative genotype calling approach in the GSMapper software (HCDiff):

- We identified up to 14,284 coding variants and small in/dels per individual exome.
- Up to 1,679 of these were putative novel polymorphisms.
- We identified up to 624 novel nonsynonymous SNPs and 54 SNPs were with +/- 2bp of an exon boundary per individual.

<table>
<thead>
<tr>
<th>Results for individual 10039:</th>
<th>KNOWN VARIANTS</th>
<th>12605</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Synonymous</td>
<td></td>
<td>5328</td>
</tr>
<tr>
<td>indel</td>
<td></td>
<td>109</td>
</tr>
<tr>
<td>SNP</td>
<td></td>
<td>5219</td>
</tr>
<tr>
<td>Synonymous</td>
<td></td>
<td>7277</td>
</tr>
<tr>
<td>indel</td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>SNP</td>
<td></td>
<td>7232</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NOVEL VARIANTS</th>
<th>1679</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Synonymous</td>
<td>1050</td>
</tr>
<tr>
<td>indel</td>
<td>426</td>
</tr>
<tr>
<td>SNP</td>
<td>624</td>
</tr>
<tr>
<td>Synonymous</td>
<td>629</td>
</tr>
<tr>
<td>indel</td>
<td>99</td>
</tr>
<tr>
<td>SNP</td>
<td>530</td>
</tr>
</tbody>
</table>

Total 14284
False negative rates across eight exomes

- Error estimations were based on 44,513 overlapping Illumina genotypes.
- Comparison of different variant calling approaches: HCDiff, ALLDiff, empirical method.
- At approx. >15x read coverage differences between tested variant calling approaches became negligible.

![Graph showing false negative genotype calls vs coverage depth for different methods: HCDiff, ALLDiff, Dynamic Calling.](image)
Sequence read distribution was compared to 44,513 Illumina homozygote (blue diamonds) and heterozygote (green triangles) genotype calls.

Symbols are placed according to the proportion of variant sequence reads per each coverage depth.

Green triangles within blue diamonds indicate partially non-concordant genotype calls.
Variant read distribution across eight exomes

I. High coverage – very confident genotype call
II. Borderline coverage for confident calls
III. Low coverage – questionable calls
Empirical optimization of calling thresholds

- Comparison to independently obtained genotypes.
- To maximize correct variant calls at lower coverage.
- Allows usage of lower covered targets.
- At ≥8x coverage 99% of SNPs were identified.

Example

<table>
<thead>
<tr>
<th>Coverage depth</th>
<th># of markers</th>
<th>Genotype calling threshold in % variant reads</th>
<th>% correct seq calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>62</td>
<td>70</td>
<td>96.8</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
<td>75</td>
<td>96.8</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
<td>80</td>
<td>96.8</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
<td>85</td>
<td>98.4</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
<td>90</td>
<td>98.4</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
<td>95</td>
<td>85.5</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
<td>100</td>
<td>85.5</td>
</tr>
</tbody>
</table>

Optimal threshold at 13x coverage is between 85% and 90% variant calls.
Based on 112,384 Illumina genotypes

- Re-sequencing of 53 randomly chosen cSNPs resulted in 100% concordance
- The empirical genotype calling approach resulted in twice as high false positive rate at lower coverage.
Advantages of a multigenerational pedigree

Mendelian inconsistencies

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Errors</th>
<th>Markers</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥5x</td>
<td>322</td>
<td>17702</td>
<td>0.01819</td>
</tr>
<tr>
<td>≥10x</td>
<td>25</td>
<td>5156</td>
<td>0.004849</td>
</tr>
<tr>
<td>≥15x</td>
<td>1</td>
<td>471</td>
<td>0.002123</td>
</tr>
</tbody>
</table>

De-novo mutations

Combining multiple related individuals into one quasi individual
Advantages of a multigenerational pedigree

Mendelian inconsistencies

De-novo mutations

- Filtering of variants in parent–offspring trios:
  47 variants tested with Sanger sequencing
  --> No de-novo event confirmed.

Combining multiple related individuals into one quasi individual
Advantages of a multigenerational pedigree

Mendelian inconsistencies

De-novo mutations

Combining multiple related individuals into one quasi individual

- We combined raw reads from three siblings for re-analysis to improve low covered targets.
- Approximately 11% additional high-confidence SNPs were identified.
- Application: Combine affected individuals in a Mendelian family.
Summary

- Exome sequencing provides a powerful means to detect variation in >90% of human exons.

- Even if “maximal” coverage can be afforded, a proportion of targets will be covered at lower levels. Array-based genotyping may be utilized to maximize variant calling.

- The study of families allows for the application of simple yet powerful Mendelian laws of inheritance to identify disease associated variation.

- A renaissance of Mendelian and family oriented genetics can be predicted, but the “needle in the haystack” problem may not be underestimated.
Acknowledgements

- Drs. Dale Hedges, Jia Huang, Margaret Pericak-Vance, Eden Martin
- Eric Powel, Cherylyn Almonte, Dr. Stuart Young, Gladys Montenegro

The study was supported by the National Institute for Neurology and Stroke (NS065712 to S.Z. and NS026630 to M.A.P-V) and the National Human Genome Research Institute (RC2HG005605 to E.M).

- Roche provided capture array and sequencing services.
- Benjamin Boese, Dan Burges, Timothy Harkins, Xinmin Zhang

- We have no financial conflicts of interest to disclose.
Exomes: Sequencing and Analysis

Pauline Ng, PhD
Assistant Professor
J. Craig Venter Institute
Gene Variation: Low-Hanging Fruit

- Many Mendelian disease mutations alter protein sequence

- To find the “functional” mutations
  - Allele frequency differences
  - Functional characterization
An Individual’s Exome

In Venter’s exome

~10,400 nonsynonymous SNVs

~700 coding indels

Build bioinformatic tools to help scientific community with functional annotation
An Individual’s Exome

In Venter’s exome

~10,400 nonsynonymous SNVs

~700 coding indels

Build bioinformatic tools to help with functional annotation

Ng, et al. PLoS Genetics 2008
What % of nsSNVs affect protein function (and can be involved in disease)?

SIFT prediction algorithm

nsSNV

Unaffected

Affected
Conserved region/or severe amino acid change

No disease “Normal”

Disease

http://sift.jcvi.org/ where anyone can submit their variants

PolyPhen, SNPs3D, MAPP
Patterns of Selection Against NsSNVs

- 14% of All NsSNVs affect protein function by SIFT.
- 12% of Common SNVs affect protein function by SIFT.
- 27% of Rare SNVs affect protein function by SIFT.
- 8% of Homozygous SNVs affect protein function by SIFT.
- 19% of Heterozygous SNVs affect protein function by SIFT.

Comparison:
- Rare >> Common
- Heterozygous >> Homozygous
An Individual’s Exome

In Venter’s exome

~10,400 nonsynonymous SNVs

~700 coding indels
50% of indels divisible by 3
There are many 1 bp indels

3 base pairs $\rightarrow$ 1 aa (In-frame)
Indels occur at beginning and end of proteins

Most of protein translated.
Indels at N-terminus can be rescued by downstream starts.
http://sift.jcvi.org

chr1, 148797595, 1, G/C
chr12, 155150158, -1, G/A
chr13, 148750264, 1, G/A
chr4, 177829363, 1, G/C
chr5, 205818112,-1, G/A

Prateek Kumar
### SIFT nsSNV Output

<table>
<thead>
<tr>
<th>Variant</th>
<th>Gene</th>
<th>Amino Acid Change</th>
<th>dbSNP</th>
<th>SIFT Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1, 148797595, 1, G/C</td>
<td>GBE</td>
<td>R190G</td>
<td>rs2229515</td>
<td>DAMAGING</td>
</tr>
<tr>
<td>chr12, 155150158, -1, G/A</td>
<td>DYNC2</td>
<td>I230L</td>
<td>novel</td>
<td>TOLERATED</td>
</tr>
<tr>
<td>chr13, 148750264, 1, G/A</td>
<td>ORC3L</td>
<td>F33S</td>
<td>rs23073289</td>
<td>TOLERATED</td>
</tr>
<tr>
<td>chr4, 177829363, 1, G/C</td>
<td>PES1</td>
<td>V217I</td>
<td>rs42942</td>
<td>TOLERATED</td>
</tr>
<tr>
<td>chr5, 205818112,-1, G/A</td>
<td>C10orf2</td>
<td>M31I</td>
<td>novel</td>
<td>DAMAGING</td>
</tr>
</tbody>
</table>

10,000 single nucleotide variants: ~ 10 minutes on the SIFT website
## SIFT Indel Website Output

<table>
<thead>
<tr>
<th>Variant</th>
<th>Gene</th>
<th>Protein Change</th>
<th>Protein Location</th>
<th>Nonsense-Mediated Decay?</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr3,81780819,81780819,</td>
<td>GBE</td>
<td>LKIR-&gt;LKsIR</td>
<td>50%</td>
<td>-</td>
</tr>
<tr>
<td>-1,AGT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2,43881516,43881516,</td>
<td>DYNC2</td>
<td>FRAMESHIFT</td>
<td>3%</td>
<td>Yes</td>
</tr>
<tr>
<td>1,AGTC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2,43857513,43857517,</td>
<td>ORC3L</td>
<td>FRAMESHIFT</td>
<td>92%</td>
<td>No</td>
</tr>
<tr>
<td>1,/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1,000 indels in 10 minutes
Next gen. exome sequencing versus Sanger sequencing

Capture exome with Nimblegen microarray → 454 sequencing → Coding variants

Samuel Levy & Ewen Kirkness
Next gen. exome sequencing versus Sanger sequencing

Capture exome with Nimblegen microarray

454 sequencing

Coding variants

Sanger ABI 3730XL

“gold standard”

assembled genome

compare

Nimblegen/454 Sequence Coverage

~90% exome covered at ≥ 10x

50x average

J. Craig Venter Institute
~90% exome covered at ≥ 4x average

Nimblegen/454 Sequence Coverage

% Capture Target Exome

0 10 20 30 40 50 60 70 80 90 100

Normalized Coverage

0 0.2 0.4 0.6 0.8 1

20x average
SNP overlap between 454 and Sanger

15.6X

454

2,522

11,530

Sanger

3,118

79% of Sanger SNPs found
SNP overlap between 454 and Sanger

- 15.6X
- 11,530 SNPs
- 3,118 SNPs
- 2,522 SNPs
- 79% of Sanger SNPs found

- Sanger has known false negative rate
- Will verify with SOLiD data
Acknowledgments

**SIFT Website**
- Prateek Kumar
- Lakshmi Radhakrishnan
- Funding: NHGRI

**454 Analysis**
- Sam Levy
- Ewen Kirkness
- Bob Strausberg

**Computational and Sequencing Support**
- Nelson Axelrod
- Kelvin Li
- Tim Stockwell
- Hue Vuong
- Brian Walenz
- Karen Beeson
- Dana Busam
- Sana Scherbakova
- Yu-Hui Rogers

**Life Sciences**
- Benjamin Boese
- Pascal Bouffard
- Zdenek Markovic
- Antonio Goncalves
- Lewyn Li
- Qin Zhao

**Roche**
- Timothy Harkins

J. Craig Venter
INSTITUTE
Look out for more webinars in the series at:

www.sciencemag.org/webinar

To provide feedback on this webinar, please e-mail your comments to webinar@aaas.org

For related information on this webinar topic, go to:

www.454.com