Integration of multiple datasets in head and neck cancer

Analysis of transcriptome and exome sequencing with methylation array data

David I Smith, Ph.D.
Professor
Chairman of the Technology Assessment Group
Center for Individualized Medicine
Mayo Clinic
Evolution of Instrument Performance

Sequencing Run Parameters

- Run format: 2x150 bp
- Output full run: 1.13Tb
- Output per day: 81Gb
- Data passing filter: 88.9%

Graph showing Gb/run from 2006 to 2011 with data points for GA and HiSeq 2000.
Uses of Next Generation Sequencing

Whole genome sequencing- requires 100+ gigabases of total sequencing (probably more for cancer genomes)

Whole exome sequencing- requires enrichment of the exome and then deep sequencing of the resulting 50 Mbs

Transcriptome sequencing- what is actively being transcribed in the cells of interest

Methylation sequencing- characterizing the epigenome, or utilize the Illumina 450K methylation arrays
Head and Neck Cancer

- Oropharynx
  - Base of Tongue
  - Tonsils

- Oral Cavity
  - Hard Palate
  - Gums
  - Lips
  - Tongue 2/3

- Larynx

- Nasopharynx

- Sinus

- Nasal cavity

- Salivary glands

- Trachea
Head and Neck cancer

- 6th most common cancer
- Epithelial Cancers (Carcinomas)
  - Squamous Cell Carcinoma
  - Adenocarcinoma
  - Thyroid
- Mesenchymal
  - (Sarcomas)
Risk Factors

- Smoking and drinking
  - 6-7\textsuperscript{th} decade of life, prolonged exposure
  - All sites

- HPV (16,18)
  - Oropharyngeal SCC (30-90\% of patients)
  - Younger patients (<50 years)
  - Lack traditional risk factors
  - Chemo/Radiation Sensitive
  - Lacking strong evidence of consistent viral integration
Head and Neck Function

- Speech/Communication
- Swallowing
- Smell
- Taste
- Vision
- Cosmesis
- Hearing
Treatment

• Overall mortality ~ 60%
  – Depends on Patient, Tumor, and Treatment

• Treatment modalities
  – Surgery: Primary vs Salvage
  – Non operative therapy
    • Radiation: Primary vs adjunct vs salvage
    • Chemotherapy: adjunctive
      – Neoadjuvant (induction), adjuvant (concomitant)
Treatment modality

• Surgery: $30,000-50,000
• Radiation: $100,000-300,000
  – Outcomes roughly equivalent
• Chemotherapy: $100,000-500,000
  – 4% survival benefit in select patients
  – 4% of patients die from treatment
  – Financial incentive to treat with Radiation/Chemo?
HPV not just in the cervix/cervical cancer

In cervical cancer HPV long-term infection coupled with eventual integration appears to be the trigger causing cervical cancer.

HPV also found in some head and neck cancers, anal cancer, vulva and vaginal cancer, and cancer of the penis.

Role of HPV in these other sites is less well understood, however.
HPV oncogenes

- E6 and E7 genes are consistently expressed in HPV infected cells
- HPV E6 decreases p53 expression
- HPV E7 gene leads to loss of pRB cell cycle arrest
- pRb negatively feeds back on p16
- When pRb is down regulated by HPV E7, p16 is over expressed
Detecting HPV in head and neck cancer

- In situ hybridization for HPV sequences
- PCR for HPV DNA
- Immunohistochemistry for p16 expression (which is elevated in HPV+ cancers)
- Detection of transcriptionally active infections by measuring RNA transcripts, or protein for the oncoproteins E6 and E7
HPV Detection by PCR

• More Consistent results from Fresh Frozen Tissues versus Paraffin Embedded Samples
HPV Detection by IHC

A: H&E Stain
B: IHC p16\(^{\text{INK4A}}\)
C: IHC p16\(^{\text{INK4A}}\) (40X)
HPV E6 and E7 Expression by Q-RT PCR

• Not routinely done
• Difficulty in doing this out of paraffin
• New technologies to do this- Nanostring, Fluidigm, or Advanced Cell Diagnostics (RNAscope)
HPV Infection Rates

• High rates of HPV positivity in BTT (oropharyngeal) tumors (26-91%), Is this because HPV rates differ in different populations or is this a technical problem?

• Low rates of HPV positivity in Oral Cavity tumors (3-11%)

• Oropharyngeal HPV+
  – HPV16,18 most common, other sub-types?
In figure 3a, the survival probability for patients with p16-positive tumors (n=26, 10 deaths) is compared with those with p16-negative tumors (n=31, 21 deaths). The p-value for this comparison is 0.067 with surgery +/- RT. In figure 3b, the survival probability for patients with p16-positive tumors (n=8, 6 deaths) is compared with those with p16-negative tumors (n=17, 16 deaths). The p-value for this comparison is 0.096 with RT or RCT.
Two distinct entities of oropharyngeal cancer?

- HPV minus – generally older smokers and drinkers. Much worse clinical outcome due to more DNA damage
- HPV positive- contains those individuals who are both younger and many are non-drinkers/non-smokers. Less DNA damage, hence possibly more treatable?
- Not clean groupings as there are smokers/drinkers that are also HPV positive
Which factors are most significant for the development of disease?
Application of Massively Parallel DNA sequencing and related technologies

mRNA-Seq / nCounter (or other) validation

Exome sequencing

Methylation arrays

Exome (8)

Methylome (24)

Transcriptome (18)

3

4

9
Application of Massively Parallel DNA sequencing and related technologies

mRNA-Seq / nCounter validation

Exome sequencing

Methylation arrays

Clinical Application - Tool Development

Discovery - Experimental

Patient knowledge of technologies - expectations
Clinical Applications

- Supplement current technologies for diagnosis (detection of HPV)
- Develop markers for patient stratification based on most significant risk factors
- Develop tools for early detection of primary disease and recurrence
Experimental Interests

- Transcript dysregulation
  - Risk factors, sites, disease behavior
- Non-coding transcripts
- Fusion transcripts
- Methylation and mutations – define mechanisms of genome dysregulation
mRNA-Seq Transcriptional profiling of Oropharyngeal SCC

• 18 Tumor and Patient Matched Normal Tissue Samples
• Illumina GAIIx (12 samples) (oligo dT-primed) and SOLiD (6 samples) (Ribominus removal of rRNA species)
• ~65 million reads per sample (200 million for SOLiD RNAseq)
• Data analysis –
  – Genesifter, Geospiza, Seattle WA
**RNAseq on T/N Pairs**

- **Illumina GA IIx or SOLiD 4**

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Illumina/SOLiD</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>M</td>
<td>Illumina</td>
</tr>
<tr>
<td>73</td>
<td>M</td>
<td>Illumina</td>
</tr>
<tr>
<td>54</td>
<td>M</td>
<td>Illumina</td>
</tr>
<tr>
<td>65</td>
<td>F</td>
<td>Illumina</td>
</tr>
<tr>
<td>48</td>
<td>M</td>
<td>SOLiD</td>
</tr>
<tr>
<td>48</td>
<td>M</td>
<td>Illumina</td>
</tr>
<tr>
<td>38</td>
<td>M</td>
<td>Illumina</td>
</tr>
<tr>
<td>64</td>
<td>F</td>
<td>Illumina</td>
</tr>
<tr>
<td>66</td>
<td>M</td>
<td>Illumina</td>
</tr>
<tr>
<td>60</td>
<td>F</td>
<td>Illumina</td>
</tr>
<tr>
<td>69</td>
<td>M</td>
<td>SOLiD</td>
</tr>
<tr>
<td>62</td>
<td>M</td>
<td>SOLiD</td>
</tr>
<tr>
<td>46</td>
<td>M</td>
<td>Illumina</td>
</tr>
<tr>
<td>49</td>
<td>F</td>
<td>Illumina</td>
</tr>
<tr>
<td>73</td>
<td>M</td>
<td>Illumina</td>
</tr>
<tr>
<td>48</td>
<td>M</td>
<td>SOLiD</td>
</tr>
<tr>
<td>55</td>
<td>M</td>
<td>SOLiD</td>
</tr>
<tr>
<td>54</td>
<td>F</td>
<td>SOLiD</td>
</tr>
</tbody>
</table>
Develop markers for patient stratification based on most significant risk factors

- Patients present with overlapping risk factors
- Patient self-reporting of tobacco exposure is often unreliable
- Lack of mechanism for role of HPV
Global Gene Expression mRNA-Seq patients grouped by smoking status
Analysis of transcriptional profiling

• Differentially regulated genes
  – HOX, MMPs, T cell signaling, Immune responsive gene targets

• Pathway analysis (Genesifter)
  – p53 signaling pathway
  – Is there a difference between groups divided by risk factor

Regulatory Non-coding Transcripts
Transcriptome Analysis: Differential T/N

PSI SIGNALING PATHWAY

- DNA damage
- Hypoxia
- Nutrient deprivation
- DNA damage
- Oncogene activation (e.g., MYC, ERK, Ras, BCR-ABL)

SFN

Target genes:
- p53
- p21
- Cyclin E
- CDK4/6
- p16
- MDM2
- MDMX

Response:
- Cell cycle arrest
- Apoptosis

Inhibition of angiogenesis and metastasis
DNA repair and damage prevention
Inhibition of IGF-1/PI3K pathway
Exposure mediated secretion
p53 negative feedback
Cellular senescence
p53 Signaling

Stress signals

- γ-irradiation
- UV
- Genotoxic drugs
- Nutrition deprivation
- Heat/cold shock

DNA damage

- Oncogene activation (Such as MYC, E2F1, Ras, BCR-ABL)

ATM
ATR

CHK2
CHK1

Hypoxia
Nitric oxide

p14ARF
MDM2
p53
CDKN2A
p16

MDM-X

Cell cycle
Validation Experiments

- Small sample size restricts analysis
- Validate in larger number of tumor/normal pairs
- Nanostring nCounter digital counter
  - 96 gene codesets
  - 44 sample pairs
  - 100ng input RNA
  - Compare fresh frozen to FFPE
nCounter Codesets

- Capture Probe
- Reporter Probe
- Target Probe Complex
- Solution Phase Hybridization
Excess Probes are Removed

Immobilize Complexes
<table>
<thead>
<tr>
<th>Color Code</th>
<th>Counts</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Color Code 1" /></td>
<td>3</td>
<td>XLSA</td>
</tr>
<tr>
<td><img src="image" alt="Color Code 2" /></td>
<td>2</td>
<td>FOX5</td>
</tr>
<tr>
<td><img src="image" alt="Color Code 3" /></td>
<td>1</td>
<td>INSULIN</td>
</tr>
</tbody>
</table>
nCounter HPV gene target analysis

• Evaluate expression patterns of genes with implicated correlation to HPV infection in head and neck cancer

• Group patients based on HPV16- E7 transcriptional levels
Average linkage cluster analysis
HPV16- E7 by qPCR

<table>
<thead>
<tr>
<th>HPV status grouping</th>
<th>Patient</th>
<th>HPV16 E6</th>
<th>HPV16 E7</th>
<th>Subsite</th>
<th>Age</th>
<th>Smoking States</th>
<th>T stage</th>
<th>N Stage</th>
<th>M Stage</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative / Latent E7</td>
<td>1</td>
<td>2.41E-06</td>
<td>2.05E-06</td>
<td>Tonsil</td>
<td>73</td>
<td>Never</td>
<td>2</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2.55E-05</td>
<td>0.17E-05</td>
<td>BOT</td>
<td>59</td>
<td>Former</td>
<td>3</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.006139</td>
<td>0.00162</td>
<td>Tonsil</td>
<td>80</td>
<td>Never</td>
<td>1</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.000162</td>
<td>0.001197</td>
<td>Tonsil</td>
<td>90</td>
<td>Former</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.000022</td>
<td>0.001202</td>
<td>Tonsil</td>
<td>49</td>
<td>Current</td>
<td>1</td>
<td>2b</td>
<td>0</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.000133</td>
<td>0.000212</td>
<td>Tonsil</td>
<td>45</td>
<td>Never</td>
<td>4a</td>
<td>2b</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.000472</td>
<td>0.000992</td>
<td>Tonsil</td>
<td>47</td>
<td>Former</td>
<td>4a</td>
<td>2c</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.000474</td>
<td>0.000601</td>
<td>Tonsil</td>
<td>73</td>
<td>Former</td>
<td>3</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.000438</td>
<td>0.001502</td>
<td>Tonsil</td>
<td>81</td>
<td>Never</td>
<td>1</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Low</td>
<td>11</td>
<td>0.044833</td>
<td>0.040049</td>
<td>Tonsil</td>
<td>50</td>
<td>Former</td>
<td>2</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.006586</td>
<td>0.100564</td>
<td>BOT</td>
<td>73</td>
<td>Never</td>
<td>3</td>
<td>2b</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>0.006534</td>
<td>0.153598</td>
<td>BOT</td>
<td>75</td>
<td>Former</td>
<td>2</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>0.006307</td>
<td>0.153561</td>
<td>BOT</td>
<td>62</td>
<td>Never</td>
<td>1</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.116239</td>
<td>0.211118</td>
<td>Tonsil</td>
<td>83</td>
<td>Former</td>
<td>3</td>
<td>2b</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.123287</td>
<td>0.267703</td>
<td>Tonsil</td>
<td>48</td>
<td>Never</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>0.123298</td>
<td>0.267728</td>
<td>Tonsil</td>
<td>49</td>
<td>Current</td>
<td>3</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0.247345</td>
<td>0.256892</td>
<td>BOT</td>
<td>30</td>
<td>Never</td>
<td>2</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>0.368558</td>
<td>0.330641</td>
<td>Tonsil</td>
<td>83</td>
<td>Never</td>
<td>2</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>High</td>
<td>20</td>
<td>0.476037</td>
<td>0.470712</td>
<td>BOT</td>
<td>57</td>
<td>Current</td>
<td>2</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0.644739</td>
<td>0.541167</td>
<td>BOT</td>
<td>55</td>
<td>Never</td>
<td>1</td>
<td>2b</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>0.081882</td>
<td>1.425063</td>
<td>BOT</td>
<td>57</td>
<td>Never</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>1.680543</td>
<td>1.774166</td>
<td>BOT</td>
<td>62</td>
<td>Former</td>
<td>2</td>
<td>2c</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>0.098721</td>
<td>2.126687</td>
<td>BOT</td>
<td>57</td>
<td>Current</td>
<td>2</td>
<td>2c</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

Negative / Latent E7

Low E7

High E7
<table>
<thead>
<tr>
<th>Large Genes</th>
<th>NonCoding Transcripts</th>
<th>p53</th>
<th>MMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANK2</td>
<td>CDKN2BAS</td>
<td>CHK2</td>
<td>MMP11</td>
</tr>
<tr>
<td>BAI3</td>
<td>HOTAIR</td>
<td>CHK1</td>
<td>MMP3</td>
</tr>
<tr>
<td>DLG2</td>
<td>UCA1</td>
<td>TP53</td>
<td>MMP14</td>
</tr>
<tr>
<td>PTPRG</td>
<td>GAS5</td>
<td>MDM2</td>
<td>MMP1</td>
</tr>
<tr>
<td>ERBB4</td>
<td>HULC</td>
<td>CDKN2A</td>
<td>MMP13</td>
</tr>
<tr>
<td>FGF14</td>
<td>KCNO1OT1</td>
<td>ATR</td>
<td>MMP12</td>
</tr>
<tr>
<td>RGS6</td>
<td>PCGEM1</td>
<td>MYC</td>
<td>Other</td>
</tr>
<tr>
<td>CTNN1D2</td>
<td>DLEU2</td>
<td>E2F1</td>
<td>RFWD2</td>
</tr>
<tr>
<td>RYR3</td>
<td>LSINCT2</td>
<td>NRAS</td>
<td>PPM1D</td>
</tr>
<tr>
<td>A2BP1</td>
<td>LSINCT3</td>
<td>AKT1</td>
<td>THBS1</td>
</tr>
<tr>
<td>CA10</td>
<td>LSINCT5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELMO1</td>
<td>LSINCT10</td>
<td>ECM</td>
<td>CYCS</td>
</tr>
<tr>
<td>HDAC9</td>
<td>LSINCT11</td>
<td>VTN</td>
<td>PERP</td>
</tr>
<tr>
<td>HS6ST3</td>
<td></td>
<td>VWF</td>
<td>SRC</td>
</tr>
<tr>
<td>KCNMA1</td>
<td>HOX Genes</td>
<td></td>
<td>PRKCG</td>
</tr>
<tr>
<td>MYT1L</td>
<td>HOXD13</td>
<td>IBSP</td>
<td>PTK2</td>
</tr>
<tr>
<td>NRXN1</td>
<td>HOXD11</td>
<td>THBS</td>
<td>PIK3CA</td>
</tr>
<tr>
<td>CTNN1A2</td>
<td>HOXC13</td>
<td>ITGAV</td>
<td>ITGB6</td>
</tr>
<tr>
<td>DCC</td>
<td>HOXA9</td>
<td>ITGA2B</td>
<td>EGFR</td>
</tr>
<tr>
<td>GRIK2</td>
<td></td>
<td>ITGB5</td>
<td>CDH1</td>
</tr>
<tr>
<td>HPSE2</td>
<td></td>
<td>ITGB8</td>
<td>CTNNBP1</td>
</tr>
<tr>
<td>NEBL</td>
<td></td>
<td>ITGB4</td>
<td></td>
</tr>
<tr>
<td>NLGN1</td>
<td></td>
<td>TNR</td>
<td></td>
</tr>
<tr>
<td>PDE4D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRKG1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Human papillomavirus type 16
Human papillomavirus type 16
Human papillomavirus type 16
Human papillomavirus - 16
CMV (HHV-5)
CMV (HHV-5)
EBV (HHV-4)
EBV (HHV-4)
Results

- Did not work out of paraffin at all
- Appears to work for more abundantly expressed transcripts, not too good for low abundance
- Worked great for HPV E6 and E7
ATR ataxia telangiectasia and Rad3 related
CHEK2
Variation in expression of Current Smoking groups
Linking Expression of FOXM1, CEP55 and HELLS to Tumorigenesis in Oropharyngeal Squamous Cell Carcinoma

Jeffrey R. Janus, MD; Rebecca R. Laborde, PhD; Alexandra J. Greenberg; Vivian W. Wang, PhD; Wei Wei, PhD; Anna Trier; Steven Olsen, MD; Eric J. Moore, MD; Kerry D. Olsen, MD; David I. Smith, PhD
What else is altered in cancer?

- Genome wide changes in methylation
- Methylation changes may be a better direct indicator of environmental exposures
- How to measure methylation? Whole genome sequencing before and after bisulfite modification is one expensive way
- Alternative is a $450 array that measures 450,000 methylation sites in the genome. Downside- only a small fraction of all methylation sites
<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>M</td>
</tr>
<tr>
<td>56</td>
<td>M</td>
</tr>
<tr>
<td>73</td>
<td>M</td>
</tr>
<tr>
<td>56</td>
<td>M</td>
</tr>
<tr>
<td>65</td>
<td>M</td>
</tr>
<tr>
<td>41</td>
<td>M</td>
</tr>
<tr>
<td>57</td>
<td>M</td>
</tr>
<tr>
<td>50</td>
<td>M</td>
</tr>
<tr>
<td>64</td>
<td>F</td>
</tr>
<tr>
<td>48</td>
<td>M</td>
</tr>
<tr>
<td>66</td>
<td>M</td>
</tr>
<tr>
<td>69</td>
<td>M</td>
</tr>
<tr>
<td>62</td>
<td>M</td>
</tr>
<tr>
<td>56</td>
<td>M</td>
</tr>
<tr>
<td>76</td>
<td>M</td>
</tr>
<tr>
<td>59</td>
<td>M</td>
</tr>
<tr>
<td>46</td>
<td>M</td>
</tr>
<tr>
<td>73</td>
<td>M</td>
</tr>
<tr>
<td>55</td>
<td>M</td>
</tr>
<tr>
<td>54</td>
<td>F</td>
</tr>
<tr>
<td>51</td>
<td>M</td>
</tr>
<tr>
<td>58</td>
<td>M</td>
</tr>
<tr>
<td>48</td>
<td>M</td>
</tr>
<tr>
<td>37</td>
<td>M</td>
</tr>
</tbody>
</table>
Background

- Methylation: addition of methyl group to a substrate or the substitution of an atom or group by a methyl group.
- DNA methylation at CpG sites
  - conversion of the cytosine to 5-methylcytosine
  - Methylation in promoter region: Inverse relationship with transcriptional activity
Methylation

- Identify differentially methylated cg sites
  - SAM:
    - We chose a delta with which the median FDR among all cg sites is 0.01.
  - Limma:
    - Adjusted P-value < 0.01.
- Overlap of limma result and SAM result: 11,383 cg sites. 49,047 are hyper-methylated, 64,806 are hypo-methylated.
Similar patterns of variability when comparing transcriptome analysis and hypermethylation of current smoking patients.
Hypermethylation of specific gene segments
Violin plots for M_tumor - M_normal

TSS200

Body
HOXA9 Hypermethylation at TSS200

Fold change of methylation in tumor compared to normal tissue

- nsm
- psm
- csm
Correlation Between Two Datasets

- 7 patients have both methylome and transcriptome data.

- Spearman correlation:
  - 14 pairs of M-value and expression value
  - cg site and the gene it lies on
  - Hyper-methylation with down-regulation; hypo-methylation with up-regulation.
Two Lists of Genes Chosen

- Genes that are differentially methylated when comparing tumor to normal and have corresponding expression change
- Genes that are differentially methylated when comparing the smokers to the non-smokers and ex-smokers
Validation

- Limited resources (both financial and in terms of tissue material)
- Fluidigm- Nanofluidics solution that can measure expression of 96 genes (or 48 in duplicate) against 96 samples
- Requires 5 nl per individual reaction
- Both cost and tissue effective- but you still need to purchase TaqMan probes for each gene
Whole Exome Sequencing

• Instead of sequencing the entire 6 Gb genome, why not “pull down” the 38 Mb exome?
• Much less sequencing and can sequence the exome at depth (100+ coverage)
SureSelect™ Target Enrichment System: Workflow

- Genomic Sample (Set of chromosomes)
- NGS Kit
- SureSelect™ Target Enrichment System Capture Process
  - Genomic Sample (Prepped)
  - SureSelect Hybridization Buffer
  - SureSelect Biotinylated RNA Library "Baits"
  - Hybridization
  - Streptavidin-Coated Magnetic Beads
  - Unbound Fraction Discarded
  - Wash Beads and Digest RNA
  - Bead Capture
  - Amplify

Agilent Technologies
SureSelect™ Target Enrichment System
Exome Sequencing

- Illumina TruSeq Library and Exome Enrichment kit

<table>
<thead>
<tr>
<th>Sample</th>
<th>Age</th>
<th>Smoking</th>
<th>HPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54</td>
<td>Current</td>
<td>Neg</td>
</tr>
<tr>
<td>2</td>
<td>46</td>
<td>Current</td>
<td>Neg</td>
</tr>
<tr>
<td>3</td>
<td>56</td>
<td>Current</td>
<td>Neg</td>
</tr>
<tr>
<td>4</td>
<td>69</td>
<td>Current</td>
<td>Neg</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>Never</td>
<td>Positive</td>
</tr>
<tr>
<td>6</td>
<td>57</td>
<td>Never</td>
<td>Positive</td>
</tr>
<tr>
<td>7</td>
<td>48</td>
<td>Never</td>
<td>Positive</td>
</tr>
<tr>
<td>8</td>
<td>54</td>
<td>Never</td>
<td>Positive</td>
</tr>
</tbody>
</table>
### Variant Summary

<table>
<thead>
<tr>
<th></th>
<th>Smokers</th>
<th></th>
<th></th>
<th></th>
<th>Non-smokers</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>463</td>
<td>527</td>
<td>691</td>
<td>718</td>
<td>663</td>
<td>687</td>
<td>716</td>
<td>721</td>
</tr>
<tr>
<td>Total SNVs</td>
<td>117,902</td>
<td>144,643</td>
<td>142,491</td>
<td>133,823</td>
<td>141,016</td>
<td>114,960</td>
<td>144,270</td>
<td>141,163</td>
</tr>
<tr>
<td>Common SNVs</td>
<td>114,195</td>
<td>142,578</td>
<td>140,236</td>
<td>124,940</td>
<td>136,549</td>
<td>106,287</td>
<td>141,058</td>
<td>137,453</td>
</tr>
<tr>
<td>Tumor-specific variants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3,707</td>
<td>2,065</td>
<td>2,255</td>
<td>8,883</td>
<td>4,467</td>
<td>8,673</td>
<td>3,212</td>
<td>3,710</td>
</tr>
<tr>
<td>non-synonymous</td>
<td>179</td>
<td>69</td>
<td>72</td>
<td>345</td>
<td>171</td>
<td>414</td>
<td>122</td>
<td>191</td>
</tr>
<tr>
<td>non (novel)</td>
<td>81</td>
<td>25</td>
<td>25</td>
<td>71</td>
<td>48</td>
<td>106</td>
<td>37</td>
<td>65</td>
</tr>
<tr>
<td>% novel</td>
<td>45.3%</td>
<td>36.2%</td>
<td>34.7%</td>
<td>20.6%</td>
<td>28.1%</td>
<td>25.6%</td>
<td>30.3%</td>
<td>34.0%</td>
</tr>
<tr>
<td>Tumor-specific mutations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,000</td>
<td>519</td>
<td>3,350</td>
<td>824</td>
<td>930</td>
<td>590</td>
<td>409</td>
<td>662</td>
</tr>
<tr>
<td>non-synonymous</td>
<td>109</td>
<td>34</td>
<td>34</td>
<td>59</td>
<td>74</td>
<td>34</td>
<td>43</td>
<td>85</td>
</tr>
<tr>
<td>non (novel)</td>
<td>69</td>
<td>16</td>
<td>18</td>
<td>40</td>
<td>36</td>
<td>15</td>
<td>25</td>
<td>56</td>
</tr>
<tr>
<td>% novel</td>
<td>63.3%</td>
<td>47.1%</td>
<td>52.9%</td>
<td>67.8%</td>
<td>48.6%</td>
<td>44.1%</td>
<td>58.1%</td>
<td>65.9%</td>
</tr>
</tbody>
</table>

Tumor-specific variant = tumor sample different than normal (most common = LOH in tumor)
Tumor-specific mutation = same as reference in normal sample, variant in cancer sample

Novel = not in dbSNP132
Patient 463 (Current smoker, Low HPV)

Total: 117,902 snvs
Common to T and N: 114,195
Tumor-specific variants: 3707
Non-synonymous: 179

Enriched pathways
Patient 463 (Current smoker, Low HPV)

Total: 117,902 snvs
Common to T and N: 114,195
Tumor-specific mutations: 1000
Non-synonymous: 109
Not in dbSNP: 69

Enriched pathways
Patient 687 (Non Smoker, High HPV)

Total: 114,960 snvs
Common to T and N: 106,287
Tumor-specific variants: 8,673
Non-synonymous: 414
Not in dbSNP: 106

Enriched pathways
Patient 687 (Non Smoker, High HPV)

Total: 114,960 snvs
Common to T and N: 106,287
Tumor-specific mutations: 943
Non-synonymous: 106
Genes mutated in at least 3 out 4 patients

Non-smokers
IGSF3*
CDC27
MUC4

Smokers
CDC27
LOC100289375
MUC4*
MUC6
PRSS3
### Combining Exome-seq and RNA-Seq

<table>
<thead>
<tr>
<th>Chrom.</th>
<th>Position</th>
<th>Gene</th>
<th>Region</th>
<th>Type</th>
<th>dbSNP</th>
<th>Exon Pop</th>
<th>Low Cov Pop</th>
<th>Ref.</th>
<th>6745: Var +</th>
<th>6746: Var +</th>
<th>6747: Var +</th>
<th>6748: Var +</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>6,278,414</td>
<td>RNF207</td>
<td>CDS</td>
<td>snv</td>
<td>rs709209</td>
<td>0</td>
<td>0</td>
<td>A</td>
<td>NC</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>chr1</td>
<td>43,149,102</td>
<td>YBX1</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>R</td>
<td>NC</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>chr2</td>
<td>10,263,618</td>
<td>RRM2</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>R</td>
<td>NC</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>chr2</td>
<td>27,600,585</td>
<td>ZNF513</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>C</td>
<td>M</td>
<td>NC</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>chr2</td>
<td>219,268,052</td>
<td>CTDS1</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>A</td>
<td>R</td>
<td>NC</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>chr5</td>
<td>34,821,902</td>
<td>RAI14</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>R</td>
<td>NC</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>chr5</td>
<td>177,638,968</td>
<td>AGXT2L2</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>A</td>
<td>R</td>
<td>NC</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>chr6</td>
<td>109,787,200</td>
<td>ZBTB24</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>R</td>
<td>NC</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>chr8</td>
<td>32,463,096</td>
<td>NRG1</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>A</td>
<td>W</td>
<td>NC</td>
<td>W</td>
<td>W</td>
</tr>
<tr>
<td>chr8</td>
<td>98,731,403</td>
<td>MTDH</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>R</td>
<td>NC</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>chr11</td>
<td>104,900,433</td>
<td>CASP1</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>T</td>
<td>Y</td>
<td>NC</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>chr11</td>
<td>134,118,751</td>
<td>THYN1</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>T</td>
<td>Y</td>
<td>NC</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>chr12</td>
<td>64,521,471</td>
<td>SRGAP1</td>
<td>CDS</td>
<td>snv</td>
<td>rs15771292</td>
<td>1</td>
<td>0</td>
<td>C</td>
<td>Y</td>
<td>NC</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>chr16</td>
<td>31,470,886</td>
<td>ARMC5</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>T</td>
<td>NC</td>
<td>W</td>
<td>W</td>
<td>W</td>
</tr>
<tr>
<td>chr17</td>
<td>7,577,105</td>
<td>TP53</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>S</td>
<td>NC</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>chr17</td>
<td>40,257,981</td>
<td>DXK8</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>C</td>
<td>M</td>
<td>NC</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>chr20</td>
<td>33,511,157</td>
<td>ACSS2</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>R</td>
<td>NC</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>chr21</td>
<td>43,708,079</td>
<td>ABCG1</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>S</td>
<td>NC</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>chr1</td>
<td>12,854,414</td>
<td>PRAMEF1</td>
<td>CDS</td>
<td>snv</td>
<td>rs1063769</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>ND</td>
<td>ND</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>chr1</td>
<td>12,907,781</td>
<td>HNRPCL1</td>
<td>CDS</td>
<td>snv</td>
<td>rs1737105, rs26443196</td>
<td>0</td>
<td>0</td>
<td>T</td>
<td>ND</td>
<td>ND</td>
<td>Y</td>
<td>NC</td>
</tr>
<tr>
<td>chr1</td>
<td>12,907,798</td>
<td>HNRPCL1</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>A</td>
<td>NC</td>
<td>NC</td>
<td>M</td>
<td>NC</td>
</tr>
<tr>
<td>chr1</td>
<td>12,907,802</td>
<td>HNRPCL1</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>C</td>
<td>T</td>
<td>NC</td>
<td>Y</td>
<td>NC</td>
</tr>
<tr>
<td>chr1</td>
<td>12,907,803</td>
<td>HNRPCL1</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>C</td>
<td>NC</td>
<td>NC</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>chr1</td>
<td>24,497,877</td>
<td>MYOM3</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>C</td>
<td>ND</td>
<td>ND</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>chr1</td>
<td>40,960,924</td>
<td>ZNF642</td>
<td>CDS</td>
<td>snv</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>NC</td>
<td>NC</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>chr1</td>
<td>144,854,581</td>
<td>PDE4D1</td>
<td>CDS</td>
<td>snv</td>
<td>rs78371650</td>
<td>0</td>
<td>0</td>
<td>T</td>
<td>NC</td>
<td>NC</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>chr1</td>
<td>152,989,626</td>
<td>FLG</td>
<td>CDS</td>
<td>snv</td>
<td>rs3126075</td>
<td>0</td>
<td>0</td>
<td>G</td>
<td>ND</td>
<td>ND</td>
<td>Y</td>
<td>NC</td>
</tr>
</tbody>
</table>

*ND = No Data*
• Correlate patterns of differential expression to
  – DNA Mutations
  – Differential Methylation
• Validate exome mutations in transcriptome
• Identify RNA editing events
• Correlate methylation patterns with differential splicing events
Moving forward

• Identify gene expression patterns of key pathways that associate with risk factor groups
  – In an individual patient, which risk factors are most significantly influencing disease development

• Compare multiple datasets to determine mechanism
  – Exome sequencing (mutations)
  – Methylation Arrays
  – Biological marker of tobacco exposure
General Conclusions

• Sequencing tools have important potential for clinical application

• Importance of building larger datasets
  – Correlate with disease outcome as they mature

• Importance of integrating datasets
  • Correlate with exposure, etiology, response to therapy

• Bottleneck involves bioinformatic challenges
Acknowledgements

Rebecca Laborde, Ph.D
Vivian Wang, Ph.D
Terra Lasho

Analysis
Yan Asmann, Ph.D
Sumit Middha
Wei Wei, Ph.D (UMR)

Geospiza
Todd Smith, Ph.D
Eric Olsen, PhD

UIUC
Jian Ma, Ph.D.
Yiyi Liu

Otolaryngology
Charles Beatty, MD
Kerry Olson, MD
Jan Kasperbauer, MD
Eric Moore, MD
Steven Olson, MD
Jeff Janus, MD

Anatomic Pathology
Joaquin Garcia, MD