Bioinformatics for 9/11

Simon Mercer
Director of Software Engineering
Gene Codes Corporation
**World Trade Center**

- Two 110 story towers
- 50,000 workers average week day
- Transportation hub - four subway lines
- Six basement levels – stores and plaza

- Collapse estimated at 130mph
- 24,000 gallons of jet fuel
- Fires burned at 1400 Fahrenheit for over three months
- 2bn pounds of rubble
- 2,749 victims, 1,549 identified (Feb 2004)
Human DNA Identification
Human Identification using DNA

- Simple Tandem Repeats (STRs)
- 13 core loci where the lengths vary based on the number of STRs
Human Identification using DNA

- Most people think of “Fingerprinting”
- 13 core loci where the lengths vary based on the number of STRs
Human Identification using DNA

- Mitochondrial DNA
- Matrilineal inheritance
- Can also sequence two, *hypervariable* regions on the mitochondrial genome
  - Not as specific as STRs but likely to be more available

HV1  HV2
Human Identification using DNA

- Single Nucleotide Polymorphisms (SNPs)
- Approx. 4M SNPs in Human nuclear DNA
- Panel of 70 SNPs from Orchid Cellmark
- Not (yet) validated for forensic casework
Kinship Analysis calculates the likelihood ratio that indicates if two people are related. In this example, as Full Siblings in the Hispanic population. (attr. Dr. George Carmody)

Probability that these are full sibs, divided by probability that two unrelated people in this population would share the alleles
Extreme Programming
Extreme Programming Overview

Team Practices
- Whole team sits together in one room
- Work at a sustainable pace
- Integrate many times per day
- Share a common vision and vocabulary
- Reflect regularly
- Converge on a coding standard

An XP Room
Dynamic pairs write all production code
Any pair can change any code

Overall Schedule
RP = Release Planning (1-3 weeks)
It. = Iteration (fixed length, 1-3 weeks)
Release to users every 1-3 months

Release Planning
Story Cards
- Index Card
- "Headline"
Customer Tests
- Tied to stories
- Automated
Customer Conversations

Estimate points for each card
Customer splits if too big
Estimate velocity

Release Plan
It. 1 It. 2 It. 3 It. 4
Release

Iteration Planning
Select stories
Yesterday’s Weather
Select as many points as were finished last iteration

Brainstorm tasks

Programming
Design Philosophy
- Design is evolutionary and emergent
- Pay as you go: Build just enough to meet today’s requirements
- Keep design as simple as possible (but no simpler)
- High quality is both a side effect and an enabling factor
- The code says everything “once and only once”

Incremental Test-First Programming
Write a test
Implement just enough to pass
Refactor

Cycle takes 5-15 minutes

Refactoring
Stepwise design improvement via safe transformations

Example: Move Method

\[
A \xrightarrow{f()} B \quad A \xrightarrow{f()} B
\]

Copyright 2002, William C. Wake. All rights reserved.
William.Wake@acm.org http://www.xp123.com
XP Timeline

- Coding started – November 5\textsuperscript{th} 2001
- First demo to OCME – December 6\textsuperscript{th} 2001
- First release delivered – December 13\textsuperscript{th} 2001
- 78 releases since then (to Feb 2004)

- 2500+ automated tests run on every code change
- 107 code development cycles
M-FISys
Mass Fatality Identification SYStem
- 22 data sources
- Data curation, not victim identification
- Aggregate formation a critical first step

(Example data used here)
Samples found to match each other on the basis of DNA (or other) evidence
21 differences between Reference -15996_618 and the consensus.

**LEGEND:**
- **Pos** - Offset in Reference -15996_618.
- **Seq** - Base in Reference -15996_618.
- **Con** - Base in the consensus.
- **Required Edit** - Change needed in Reference -15996_618 to match corresponding base in the consensus.

<table>
<thead>
<tr>
<th>Pos</th>
<th>Seq</th>
<th>Con</th>
<th>Required Edit</th>
</tr>
</thead>
<tbody>
<tr>
<td>16,037</td>
<td>A</td>
<td>N</td>
<td>Change base</td>
</tr>
<tr>
<td>16,069</td>
<td>C</td>
<td>T</td>
<td>Change base</td>
</tr>
<tr>
<td>16,093</td>
<td>T</td>
<td>C</td>
<td>Change base</td>
</tr>
<tr>
<td>16,126</td>
<td>T</td>
<td>C</td>
<td>Change base</td>
</tr>
<tr>
<td>16,240</td>
<td>A</td>
<td>N</td>
<td>Change base</td>
</tr>
<tr>
<td>16,247</td>
<td>A</td>
<td>N</td>
<td>Change base</td>
</tr>
<tr>
<td>16,261</td>
<td>C</td>
<td>T</td>
<td>Change base</td>
</tr>
<tr>
<td>16,274</td>
<td>G</td>
<td>A</td>
<td>Change base</td>
</tr>
<tr>
<td>16,355</td>
<td>C</td>
<td>T</td>
<td>Change base</td>
</tr>
<tr>
<td>16,396</td>
<td>T</td>
<td>N</td>
<td>Change base</td>
</tr>
</tbody>
</table>

**Chromatograms from Contig[00011]:**

- A fragment base #15_037, Base 21 of 27
Samples available for DNA analysis
### Personal Effect Samples

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>Item #</th>
<th>Sample Name</th>
<th># of Loci</th>
<th>mtDNA Length</th>
<th># of SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toothbrush: PE</td>
<td>Albany</td>
<td>1</td>
<td>CORE-PE-034070-01</td>
<td>16</td>
<td></td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PE-94070-01</td>
<td>16</td>
<td></td>
<td>71</td>
</tr>
</tbody>
</table>

### Kin Samples

<table>
<thead>
<tr>
<th>Type</th>
<th>Source</th>
<th>Item #</th>
<th>Sample Name</th>
<th># of Loci</th>
<th>mtDNA Length</th>
<th># of SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buccal: BM</td>
<td>Albany</td>
<td>0</td>
<td>CORE-BM-C2017 #00</td>
<td>16</td>
<td>689</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BM-02017 #00</td>
<td>16</td>
<td>689</td>
<td>71</td>
</tr>
<tr>
<td>Buccal: BD</td>
<td>Albany</td>
<td>2</td>
<td>CORE-BD-08707 #02</td>
<td>16</td>
<td></td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BD-08707 #02</td>
<td>16</td>
<td></td>
<td>71</td>
</tr>
<tr>
<td>Buccal: PR</td>
<td>Albany</td>
<td>3</td>
<td>CORE-PR-07192 #03</td>
<td>-</td>
<td></td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PR-07192 #03</td>
<td>-</td>
<td></td>
<td>71</td>
</tr>
</tbody>
</table>

- Hide Names
Family relationships of victims
<table>
<thead>
<tr>
<th>Profiles</th>
<th>VIRT-DM0111673</th>
<th>BM-01651 #01</th>
<th>BU-51601 #01</th>
<th>BU-64642 #01</th>
<th>BS-51602 #01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen</td>
<td>XY</td>
<td>XX</td>
<td>XY</td>
<td>XY</td>
<td>XY</td>
</tr>
<tr>
<td>D3S1358</td>
<td>16</td>
<td>14/16</td>
<td>16</td>
<td>16</td>
<td>15/16</td>
</tr>
<tr>
<td>FWA</td>
<td>17/19</td>
<td>19/20</td>
<td>17/19</td>
<td>17/19</td>
<td>17</td>
</tr>
<tr>
<td>FGA</td>
<td>22/25</td>
<td>22/25</td>
<td>22/25</td>
<td>22/25</td>
<td>22</td>
</tr>
<tr>
<td>D8S1179</td>
<td>14/16</td>
<td>14</td>
<td>14/16</td>
<td>14/16</td>
<td>14/15</td>
</tr>
<tr>
<td>D21S11</td>
<td>32.2</td>
<td>32.2</td>
<td>32.2</td>
<td>32.2</td>
<td>31/32.2</td>
</tr>
<tr>
<td>D18S51</td>
<td>17/18</td>
<td>18</td>
<td>17/18</td>
<td>17/18</td>
<td>17/18</td>
</tr>
<tr>
<td>D6S818</td>
<td>12/13</td>
<td>8/13</td>
<td>12/13</td>
<td>12/13</td>
<td>12/13</td>
</tr>
<tr>
<td>D13S317</td>
<td>12/14</td>
<td>11/14</td>
<td>12/14</td>
<td>12/14</td>
<td>12/14</td>
</tr>
<tr>
<td>D7S820</td>
<td>8</td>
<td>8/9</td>
<td>8</td>
<td>8</td>
<td>8/11</td>
</tr>
<tr>
<td>D16S539</td>
<td>9/14</td>
<td>9/14</td>
<td>9/14</td>
<td>9/14</td>
<td>9/14</td>
</tr>
<tr>
<td>TH01</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6/7</td>
</tr>
<tr>
<td>TPOX</td>
<td>6/8</td>
<td>6/9</td>
<td>6/8</td>
<td>6/8</td>
<td>6</td>
</tr>
<tr>
<td>CSF1PO</td>
<td>8/12</td>
<td>8/14</td>
<td>8/12</td>
<td>8/12</td>
<td>9/12</td>
</tr>
<tr>
<td>Penta D</td>
<td>6/8</td>
<td>neg</td>
<td>neg</td>
<td>neg</td>
<td>6/9</td>
</tr>
<tr>
<td>Penta E</td>
<td>11</td>
<td>neg</td>
<td>neg</td>
<td>neg</td>
<td>11/15</td>
</tr>
</tbody>
</table>

Min LR to V: 1.4E+023, 2.7E+006, 9.9E+008, 9.9E+008, 2.0E+007
Quality control and other data curation tools