It’s all in the genes

The power of the Oracle database and Exadata in cancer research

adapted from the DOE Joint Genome Institute website
www.jgi.doe.gov
used with permission

@rjlkuipers
rkuipers@vxcompany.com
About VX Company

- What we do
  - IT-services, Oracle & Java, Managed Services
- Since
  - 1988
- Where
  - Baarn
- Number of Professionals
  - 300
- Turnover 2012
  - € 38 mln
- Certification
  - ISO 9001:2000
About me

- Business Manager Data and BI Solutions
- Datawarehouse Architect
- Business Intelligence specialist
- Master degree in Biochemistry
  - molecular biology
  - cancer genetics
Agenda

- Basic genetics
  - analyses
- Technology behind this
- What does it look like
- The next step: combining genomic data with patient data
- When both worlds meet
Set the context

BASIC GENETICS
Chromosomes
Genes

50 million base pairs

- Short stature homeo box, Y-linked
- Short stature
- Leri-Weill dyschondrosteosis
- Langer mesomelic dysplasia
- Interleukin-3 receptor, Y chromosomal
- Sex-determining region Y (testis-determining)
- Gonadal dysgenesis, XY type
- Protocadherin 11, Y-linked
- Azoospermia factors
- Male infertility due to spermatogenic failure
- Growth control, Y-chromosome influenced
- Chromodomain proteins
- Retinitis pigmentosa, Y-linked
DETERMINING THE GENETIC SEQUENCE

basic genetics
Genetic sequence

- Blood / cancer tissue
- DNA isolation
- DNA amplification
- DNA Sequencing (40x - 80x)
Genetic sequence

- approx. 5% of DNA is gene
- approx. 95% of DNA is referred to as ‘junk-DNA’

- 99% of entire DNA sequence is stable
- Genetic variations are normal
DNA (Next Generation) Sequencing
From blood-sample to DNA sequence
DNA (Next Generation) Sequencing

From blood-sample to DNA sequence

@rjlkippers
rkuipers@vxcompany.com
DNA (Next Generation) Sequencing
From blood-sample to DNA sequence

- 3 billion basepairs
DNA (Next Generation) Sequencing
From blood-sample to DNA sequence

- 3 billion basepairs
- 2 TB per sample
DNA (Next Generation) Sequencing
From blood-sample to DNA sequence

- 3 billion basepairs
- 2 TB per sample
- unique: whole genomes
Abnormal genetic variations
Abnormal genetic variations
Abnormal genetic variations

- mutations
  - inserts
  - deletes
  - substitutions

- non-functioning cells

- uncontrolled cell-growth
  - p53 protein, coded by p53 gene
  - key characteristic of cancer

- variations outside genes
Searching for the unknown
Searching for the unknown

- genetic variations $\xrightarrow{}$ normal
- genetic variations $\xleftarrow{}$ cancer
Searching for the unknown

- genetic variations $\leftrightarrow$ normal
- genetic variations $\leftrightarrow$ cancer
- better diagnoses require better analyses.
Searching for the unknown

- genetic variations ↔ normal
- genetic variations ↔ cancer

- better diagnoses require better analyses.
- Upfront (predictive) diagnoses require a lot of data and processing power.
- result: less-invasive treatment, better patient-life.
Searching for the unknown

- genetic variations ↔ normal
- genetic variations ↔ cancer

- better diagnoses require better analyses.
- Upfront (predictive) diagnoses require a lot of data and processing power.
- result: less-invasive treatment, better patient-life.

- What did we not know (yet)
  - and can be learned from
- Ultimate goal: centralized DNA library for statistical purposes

@rjkuipers
rkuiers@vxcompany.com
THE TECHNOLOGY BEHIND THIS
DNA (Next Generation) Sequencing

- 3 billion basepairs
- 2 TB per sample
- Whole genomes
Handling large volumes
Handling large volumes

- Oracle Database
  - Partitioning
  - Optimized data model
Handling large volumes

- **Oracle Database**
  - Partitioning
  - Optimized data model

- **Oracle Exadata Database Machine**
  - Optimized to run Oracle Database
  - Specific performance features
    - Smart Scans
    - Exadata Hybrid Columnar Compression
Handling large volumes

- Oracle Database
  - Partitioning
  - Optimized data model

- Oracle Exadata Database Machine
  - Optimized to run Oracle Database
  - Specific performance features
    - Smart Scans
    - Exadata Hybrid Columnar Compression

- Performance increase: 700x
Handling large volumes - database benefits

- **Datamodel V1**
  - Sample-oriented (partitioned)
  - Each base-position stored (compared to reference genome)
    - leads to 95% no-calls
  - 206 samples --> 800 GB
    - max 2,500 samples on Exadata
  - Indexes are (still) needed: Index size 5x larger than sample-size
Handling large volumes - database benefits

- Datamodel V2
  - Sample-oriented (partitioned)
  - Positions are stored as regions (buckets)
    - 1000 positions per region
  - Buckets are indexed
  - EHCC Compression
  - Reduce redundant data
    - Store allele 1 and 2 as 1 row when values are equal
  - Storage 99GB (246 samples)
    - Up to 20,000 samples

- Indexes require less space than in Datamodel V1
Exadata benefits

- Flash
- Parallel processing
- Smart Scans
- Exadata Hybrid Columnar Compression

Let’s have a look...
[oracle@dm01db01 vxone1] $ rsqplus cg/cg

SQL*Plus: Release 11.2.0.2.0 Production on Sat Apr 30 14:33:11 2011
Copyright (c) 1982, 2010, Oracle.  All rights reserved.

Connected to:
Oracle Database 11g Enterprise Edition Release 11.2.0.2.0 - 64bit Production
With the Partitioning, Real Application Clusters, Automatic Storage Management, OLAP,
Data Mining and Real Application Testing options

SQL>
### Executed tests

<table>
<thead>
<tr>
<th>Nr</th>
<th>Exadata features</th>
<th>Parallel</th>
<th>Disk type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>5</td>
<td>SS</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>6</td>
<td>SS</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>7</td>
<td>SS</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>8</td>
<td>SS</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>9</td>
<td>SS + EHCC</td>
<td>64</td>
<td>FDD</td>
</tr>
</tbody>
</table>
## Executed tests

<table>
<thead>
<tr>
<th>Nr</th>
<th>Exadata features</th>
<th>Parallel</th>
<th>Disk type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>5</td>
<td>SS</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>6</td>
<td>SS</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>7</td>
<td>SS</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>8</td>
<td>SS</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>9</td>
<td>SS + EHCC</td>
<td>64</td>
<td>FDD</td>
</tr>
</tbody>
</table>
Data Mining and Real Application Testing options

SQL> set timing on
SQL> alter session set cell_offload_processing=false;
Session altered.
Elapsed: 00:00:00.00
SQL>
# Executed tests

<table>
<thead>
<tr>
<th>Nr</th>
<th>Exadata features</th>
<th>Parallel</th>
<th>Disk type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>5</td>
<td>SS</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>6</td>
<td>SS</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>7</td>
<td>SS</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>8</td>
<td>SS</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>9</td>
<td>SS + EHCC</td>
<td>64</td>
<td>FDD</td>
</tr>
</tbody>
</table>
Data Mining and Real Application Testing options

SQL> set timing on
SQL> alter session set cell_offload_processing=false;
Session altered.
Elapsed: 00:00:00.00
SQL>
# Executed tests

<table>
<thead>
<tr>
<th>Nr</th>
<th>Exadata features</th>
<th>Parallel</th>
<th>Disk type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>5</td>
<td>SS</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>6</td>
<td>SS</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>7</td>
<td>SS</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>8</td>
<td>SS</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>9</td>
<td>SS + EHCC</td>
<td>64</td>
<td>FDD</td>
</tr>
</tbody>
</table>
Data Mining and Real Application Testing options

SQL> set timing on
SQL> alter session set cell_offload_processing=false;
Session altered.
Elapsed: 00:00:00.00
SQL>
## Executed tests

<table>
<thead>
<tr>
<th>Nr</th>
<th>Exadata features</th>
<th>Parallel</th>
<th>Disk type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>5</td>
<td>SS</td>
<td>Serial</td>
<td>HDD</td>
</tr>
<tr>
<td>6</td>
<td>SS</td>
<td>Serial</td>
<td>FDD</td>
</tr>
<tr>
<td>7</td>
<td>SS</td>
<td>64</td>
<td>HDD</td>
</tr>
<tr>
<td>8</td>
<td>SS</td>
<td>64</td>
<td>FDD</td>
</tr>
<tr>
<td>9</td>
<td>SS + EHCC</td>
<td>64</td>
<td>FDD</td>
</tr>
</tbody>
</table>
Data Mining and Real Application Testing options

SQL> set timing on
SQL> alter session set cell_offload_processing=false;

Session altered.

Elapsed: 00:00:00.00
SQL>
## Query performance

<table>
<thead>
<tr>
<th>Nr</th>
<th>Exadata features</th>
<th>Parallel</th>
<th>Disk type</th>
<th>11.2.0.1</th>
<th>11.2.0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>Serial</td>
<td>HDD</td>
<td>695</td>
<td>153</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Serial</td>
<td>FDD</td>
<td>403</td>
<td>91</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>64</td>
<td>HDD</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>64</td>
<td>FDD</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>SS</td>
<td>Serial</td>
<td>HDD</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>6</td>
<td>SS</td>
<td>Serial</td>
<td>FDD</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>7</td>
<td>SS</td>
<td>64</td>
<td>HDD</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>SS</td>
<td>64</td>
<td>FDD</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>SS + EHCC</td>
<td>64</td>
<td>FDD</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
WHAT DOES IT LOOK LIKE?
@rjkuipers
rkuipers@vxcompany.com

Saturday, 18 May 13
Why is this important?
Why is this important?

- **Speed**
  - Faster results
  - ‘No’ is found earlier
Why is this important?

- **Speed**
  - Faster results
  - ‘No’ is found earlier

- **Volume (Centralized DNA Library)**
  - Better statistical basis
  - Less-invasive treatments for patients
  - Personalized healthcare
Even more...

- Add clinical data to genomic data.
  - Patient history
  - Drug treatment history
  - Demographics
Oracle Translational Research Center (TRC)

- Oracle Healthcare Data Warehouse Foundation (OHDWF)
- Oracle Health Sciences Omics Data Bank (ODB)

**EHAD Data Integration & Validation**
- Interface tables
- ETLs from Interface tables to HDM
- Rules and metadata for data validation

**SAS**

**TRC "App Exchange"**
- Oracle Apps
  - Cohort Explorer
- Partner Apps
  - Apps
- Open Source apps
  - Command Line

**Research Systems**
- Source ETLs provided by SI partners or HSGDU consulting

@rjkuipers
rkuipers@vxcompany.com
Advanced visualizations
Summary

- Care is primary.  
  - Technology is supporting.
- Oracle offers platforms to provide better care
  - Database
  - Exadata
  - TRC
- Clinical and Genomic data are complimentary.
- Not everything is in the genes...
Q&A