Single Nucleotide Polymorphisms and Copy Number Variation

Francesca Demichelis, PhD
Dep. of Pathology and Laboratory Medicine
Institute for Computational Biomedicine
Weill Cornell Medical College
6 billion nucleotides organized in two sets of 23 chromosomes
(22 autosomes, 2 sex chromosomes)
DNA encodes 30,000 genes
Single Nucleotide Polymorphism (SNP)

Mouse

Moose

Copy Number Variants (CNV)

1 page == 1 stretch of DNA (gene)

DIVERSITY

1 copy

5 copies

F Demichelis WCMC

3/16/09
✓ Genome diversity
✓ Copy Number Variant – data evaluation
✓ SNP Panel Identification Assay
GENETIC VARIATION (1)

Point mutations
(sequence variation affecting single amino acid - Single Nucleotide Polymorphism SNP)

A - Adenine; T - Thymine; C - Cytosine; G - Guanine.

~0.1% difference in the genomes of 2 unrelated individuals
GENETIC VARIATION (1)

- Non-coding region (SNPs that are not in protein-coding regions may still have consequences for gene splicing, transcription factor binding, or the sequence of non-coding RNA).

- Coding region:
  
  o Synonymous (both forms lead to the same polypeptide sequence is termed synonymous (sometimes called a silent mutation));

  o Nonsynonymous (if a different polypeptide sequence is produced);

Variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents.

Variations in the DNA sequences of humans help understanding the genetic structure of human populations
Example 1: genetic ancestry

Nature 2008, Novembre et al: Genes mirror geography within Europe

[...The results emphasize that when mapping the genetic basis of a disease phenotype, spurious associations can arise if genetic structure is not properly accounted for. In addition, the results are relevant to the prospects of genetic ancestry testing; an individual’s DNA can be used to infer their geographic origin with surprising accuracy—often to within a few hundred kilometres...]

Principal component analysis
Example 2: disease susceptibility

Prostate cancer has strong genetic component

Contingency tables, odds ratios, Fisher test, p-value, CI
**GENETIC VARIATION (2)**

- **Insertion (unbalanced translocation)** involving 2 chromosomes
- **Balanced Translocation**
- **Variation of DNA ‘quantity’**
  - **Copy Number Variants**
    - Deletion
    - Duplication
    - Inversion

3/16/09

F Demichelis WCMC
LEXICON - CYTOGENOMICS

Global variation in copy number in the human genome

Nature 2006

Richard Redon1, Shumpei Ishikawa2,3, Karen R. Fitch4, Lars Feuk2,6, George H. Perry7, T. Daniel Andrews1, Heike Fiegler1, Michael H. Shapero1, Andrew R. Carson5,6, Wenwei Chen1, Eun Kyung Cho1, Stephanie Dallaire7, Jennifer L. Freeman1, Juan R. Gonzalez3, Monica Gratacos3, Jing Huang1, Dimitrios Kalaitzopoulos1, Daisuke Komura1, Jeffrey R. MacDonald1, Christian R. Marshall5,6, Rui Mei1, Lyndal Montgomery1, Kunihiro Nishimura1, Kohji Okamura1,6, Fan Shen1, Martin J. Somerville1, Joelle Tchinda1, Armand Valsesia1, Cara Woodwark1, Fenestane Yazd1, Junjun Zhang1, Tatiana Zerial1, Jane Zhan1, Lluis Armanco1,8...

Scherer S et al

3/16/09
Dosage variable CNVs - example

- CN of the salivary amylase gene (AMY1) correlates with salivary amylase protein level

- Individuals from high-starch diet populations have higher number of AMY1 copies

- CNV formation event maybe sufficient to specifically promote gene expression modification; thus gene copy number changes may facilitate evolutionary adaptation involving protein abundance change.

Perry GH et al, Nat Gen 2007
Why are copy number variants important/interesting?

- Hundreds of CNVs per individual. 20% potentially affecting protein-coding genes.

- CNV genesis occurs at higher rate than point mutation (1e-4/1e-6 vs 2e-8 per generation). They carry different information.

- DNA variation extension: CNVs more likely to affect coding sequence than point mutation.

- Relationship bw gene dosage and mRNA expression is basis for phenotypic traits (disease susceptibility and dietary preferences).

- CNVs tend to affect specific gene functional categories (such as environmental response related and not basic cellular processes related)

- Duplication of genetic material is common cause of protein birth (protein evolution).

? Characterization is not complete

? Mechanism behind formation of CNVs is not clear. Formation bias: non uniform distribution along chromosome.
Single Nucleotide Polymorphism (SNP)

~ 0.1% difference between the genomes of 2 unrelated individuals

Copy Number Variants (CNV)

~0.4% difference between the genomes of 2 unrelated individuals

Perry et al, Am J Hum Genet 2008 Mar;82(3):685-95
Importance of polymorphisms

• Population study/Evolution
• Phenotype/traits
• Disease risk
How to evaluate Copy Number Variants in high-throughput fashion

Arrays – aiming at high resolution AND high sensitivity

**Figure 1 | Protocol outline for two CNV detection platforms.** The experimental procedures for comparative genome hybridization on the WGTG array and comparative intensity analysis on the 500K EA platform are shown schematically (see Supplementary Methods for details), for a comparison of two male genomes (NA10851 and NA19007). The genome profile shows the log₂ ratio of copy number in these two genomes chromosome-by-chromosome. The 500K EA data are smoothed over a five-probe window. Below the genome profiles are expanded plots of chromosome 8, and a 10-Mb window containing a large duplication in NA19007 identified on both platforms (indicated by the red bracket).
SNP arrays

1) Noisier data
2) Genotype call for SNP markers

Allele specific information:

<table>
<thead>
<tr>
<th>S1_A1 : 1 copy</th>
<th>S1 : 2 copies</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1_A2 : 1 copy</td>
<td></td>
</tr>
<tr>
<td>S1_A1 : 0 copy</td>
<td>S1 : 2 copies</td>
</tr>
<tr>
<td>S1_A2 : 2 copies</td>
<td></td>
</tr>
</tbody>
</table>

3) Reference sample at data level

aCGH

1) High signal to noise ratio
2) No SNP information

3) Reference sample at experimental level

3/16/09
F Demichelis (WCMC)
A. Preprocessing (normalization)

B. Quality Control (exclusion of bad sample data based on predefined measures)

B. Evaluation of ratio of target/reference signal on a ‘marker’ basis

C. Analysis of signal along the genome (consider signal values of neighboring markers to control for noise and to define breakpoints of copy number variations – segmentation)

**Very easy in theory:** for each locus/marker $i$:

\[
\log_2(\frac{CN2_{\text{target}}}{CN2_{\text{reference}}}) = 0 \quad \text{(Normal)}
\]

\[
\log_2(\frac{CN1_{\text{target}}}{CN2_{\text{reference}}}) = -1 \quad \text{(Hemizygous del)}
\]

\[
\log_2(\frac{CN0_{\text{target}}}{CN2_{\text{reference}}}) = -\infty \quad \text{(Homozygous del)}
\]

\[
\log_2(\frac{CN3_{\text{target}}}{CN2_{\text{reference}}}) = 0.58 \quad \text{(Gain 1 copy)}
\]
To set thresholds for CN states, one needs:
- gold standard
- previous experimentally validated data;
- ad hoc experiment (FISH, PCR absolute curve)
Data Signal Processing (3) – parameter tuning

Images removed
Data Signal Processing (4) – reference model

Images removed
No perfect data / no perfect analysis

Validation of interesting variants/results by different experimental procedure

Room for improvement for statistical approaches to process CN data (analytical approaches for threshold setting, breakpoint identification, correction for reference)
Diversity makes each individual unique

- DNA test for paternity
- legal issues
- identity check of biological material

SNP Panel Identification Assay (SPIA):
a genetic-based assay for the identification of samples
Risk in cell line maintenance is human error, either by mislabeling or cross-contamination. Potential common problem (MedLine search of ‘cell line’ and ‘cancer human’ identified 96,758 studies on cancer biology.)

- Cancer genomics involves the accurate verification of sample provenance, as well as continual tracking to ensure accurate identity.
- Population studies
SNP PANEL IDENTIFICATION ASSAY (SPIA)

FINGERPRINTING

SAMPLE 1
Extract DNA

SAMPLE 2
Extract DNA

REFERENCE DB

PAIR-WISE COMPARISON

EVALUATION of GENOTYPE DISTANCE D & PROBABILISTIC TEST

DISTANCE D

TEST NOT APPLICABLE

SIMILAR
UNCERTAIN
DIFFERENT

TEST on ADDITIONAL SET of SNPs

Alternate Uses

Confirmation of Xenografts after passage
Confirmation of Passaged Cell Lines
Identification of Duplicate Samples (in silico)

Nucleic Acids Research, 2008
### PROBLEM STATEMENT

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Tumor</th>
<th>1 mismatches out of 24 SNPs (concordance 95.8%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>A A B</td>
<td>AB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AB B B</td>
<td>A A A</td>
<td>AB AB AB AB A A B</td>
</tr>
<tr>
<td>Tumor</td>
<td>A A B</td>
<td>AB A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B B A A AB AB A A B B AB AB AB AB A A B</td>
<td>1 mismatches out of 22 SNPs (concordance 95.4%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>A A B</td>
<td>AB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AB B B</td>
<td>A A A</td>
<td>AB AB AB AB A A B</td>
</tr>
<tr>
<td>Tumor</td>
<td>A A B</td>
<td>AB A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B B A A AB AB A A B B AB AB AB AB A A B</td>
<td>0 mismatches out of 12 (100%)</td>
<td></td>
</tr>
</tbody>
</table>

How many markers to use? Confidence in calling identity?

3/16/09 F Demichelis WCMC
TYPES of MISMATCHES

(CALL\textsubscript{on\_NORMAL}, CALL\textsubscript{on\_TUMOR})

LOH: \((AB,A)\) \(P(AB,A) = P(AB) \times P(A|AB)\)

GOH: \((A,AB)\) \(P(A,AB) = P(A) \times P(AB|A)\)

Doub. Mut: \((A,B)\) \(P(A,B) = P(A) \times P(B|A)\)

GENOTYPE CALL ERROR RATE (PLATFORM and ALGORITHM DEPENDENT)

Mismatches are ‘possible’ and need to be considered

Probabilities can be SNP specific and/or tissue specific

The joint probability of two events \(E1\) and \(E2\), \(P(E1,E2)\) or \(P(E1 \text{ AND } E2)\) is \(P(E1,E2) = P(E1) \times P(E2|E1)\) and \(P(E2|E1)\) is called the conditional probability of \(E2\) given \(E1\).
APPROACH

To identify the ideal SNP panel, which maximizes the probability of obtaining distinct genotype calls on different samples with ‘reasonable confidence’.

MULTI-STEP COMPUTATION to BUILD AND VALIDATE SPIA
1. definition of a genotype distance to compare samples,
2. filtering,
3. iterative procedure of training and test steps (with bootstrap) to identify best SNPs,
4. implementation of a probabilistic test (different, uncertain, or similar),
5. in silico validation on independent dataset,
6. lab validation on cell lines genotyped on independent platform (Sequenom).

DATASETS
50K genotype data of 155 cancer cell lines (CLs) derived from different organs.
SPIA panel genotype data of 93 CLs generated on on a mass spectrometer system (Sequenom).
COMPARING GENOTYPES

To count the number of loci where the two samples do not match and normalize on the total number of loci. This value is the ‘distance’ between the two samples.

This distance is proportional to the number of discordant calls.

\[
d(s_{1i}, s_{2i}) = \begin{cases} 
1 & \text{if } s_{1i} \neq s_{2i} \\
0 & \text{if } s_{1i} = s_{2i}
\end{cases}
\]

\[
D(s_1, s_2) = \frac{\sum_{i=1}^{N_{SNPs}} w_i(j) d(s_{1i}, s_{2i})}{\sum_{i=1}^{N_{SNPs}} w_i(j)}
\]

\[\nu N_{SNPs} \leq N_{SNPs}\]

\[W_i\] depends on type \(j\) (match, LOH, GOH, DM)
Pair-wise distance after SNP filtering

062906 Pair-wise comparison: 160 samples - on 5279 SNPs -

Distance D (% of discordant genotype calls)

Index of Cell Line pairs (all combinations)

3/16/09
F Demichelis WCMC
Table 1: List of cell lines which are detected to have very similar genotype profile evaluated on a set of 5.3K SNPs.

<table>
<thead>
<tr>
<th>CL1 Name</th>
<th>CL2 Name</th>
<th>D</th>
<th>% Valid Calls</th>
<th>Ho-Ho</th>
<th>Ho-Het</th>
</tr>
</thead>
<tbody>
<tr>
<td>M14</td>
<td>MDA.MB435</td>
<td>0.0747</td>
<td>0.794</td>
<td>2</td>
<td>311</td>
</tr>
<tr>
<td>MCF7</td>
<td>BT.20</td>
<td>0.0279</td>
<td>0.781</td>
<td>0</td>
<td>115</td>
</tr>
<tr>
<td>MCF7</td>
<td>KPL.1</td>
<td>0.0271</td>
<td>0.797</td>
<td>0</td>
<td>114</td>
</tr>
<tr>
<td>NCI.ADR.RES</td>
<td>OVCAR.8</td>
<td>0.0076</td>
<td>0.874</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>NCI.H460</td>
<td>H2195</td>
<td>0.0680</td>
<td>0.696</td>
<td>0</td>
<td>250</td>
</tr>
<tr>
<td>SNB.19</td>
<td>U251</td>
<td>0.0394</td>
<td>0.866</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>184A1</td>
<td>184B5</td>
<td>0.1084</td>
<td>0.978</td>
<td>37</td>
<td>523</td>
</tr>
<tr>
<td>BT.20</td>
<td>KPL.1</td>
<td>0.0308</td>
<td>0.831</td>
<td>1</td>
<td>134</td>
</tr>
<tr>
<td>H1450</td>
<td>H2141</td>
<td>0.0092</td>
<td>0.866</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td>H1450</td>
<td>H220</td>
<td>0.0000</td>
<td>0.861</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H2141</td>
<td>H220</td>
<td>0.0088</td>
<td>0.857</td>
<td>0</td>
<td>40</td>
</tr>
</tbody>
</table>
The Hardy–Weinberg principle: both allele and genotype frequencies in a population remain constant from generation to generation unless specific disturbing influences are introduced (as non-random mating, mutations, selection, limited population size, random genetic drift and gene flow).

Genetic equilibrium is an ideal state that provides a baseline to measure genetic change against.
Table 2: summary of pair-wise distances/differences varying the number of selected SNPs

<table>
<thead>
<tr>
<th>Number of SNPs</th>
<th>Set of CLs (133) used for the SNP selection process</th>
<th>Set of CLs (13) used for independent validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>~58000(*)</td>
<td>Mean $D$ (stdev) 0.3832 $D$ Min-Max 0.2653 - 0.4855</td>
<td>Mean $D$ (stdev) 0.4227 (0.0330) $D$ Min-Max 0.3613 - 0.4962</td>
</tr>
<tr>
<td>5279 (**)</td>
<td>Mean $D$ (stdev) 0.4723 (0.0328) $D$ Min-Max 0.3774 - 0.5765</td>
<td>Mean $D$ (stdev) 0.4967 (0.0337) $D$ Min-Max 0.4274 - 0.5699</td>
</tr>
<tr>
<td>80</td>
<td>Mean $D$ (stdev) 0.66 (0.06) $D$ Min-Max 0.44 - 0.86</td>
<td>Mean $D$ (stdev) 0.65 (0.06) $D$ Min-Max 0.49 – 0.78</td>
</tr>
<tr>
<td>60</td>
<td>Mean $D$ (stdev) 0.66 (0.07) $D$ Min-Max 0.37 – 0.90</td>
<td>Mean $D$ (stdev) 0.65 (0.06) $D$ Min-Max 0.50 – 0.77</td>
</tr>
<tr>
<td>40</td>
<td>Mean $D$ (stdev) 0.66 (0.09) $D$ Min-Max 0.28 – 0.94</td>
<td>Mean $D$ (stdev) 0.65 (0.08) $D$ Min-Max 0.46 – 0.85</td>
</tr>
<tr>
<td>20</td>
<td>Mean $D$ (stdev) 0.66 (0.12) $D$ Min-Max 0.20 - 1</td>
<td>Mean $D$ (stdev) 0.64 (0.11) $D$ Min-Max 0.40 – 0.90</td>
</tr>
</tbody>
</table>

LEGEND: CL = cell line; (*) set of SNPs represented on the 50K Xba chip; (**) set of filtered SNPs, used for the selection of the best SNPs.
COMPARING GENOTYPES

QUESTIONS:

1. How close (far) two samples need to be to be called ‘similar’ (‘different’)? How confident we are?

2. What is the minimum number of loci we need to make a decision?

PROBABILISTIC APPROACH

To evaluate the number of mismatches (matches) and to compare with expectations (gold standard)
PROBABILISTIC APPROACH

UNDER THE ASSUMPTION THAT
SNP calls are independent, e.g. call at locus \( i \) does not depend on
call at locus \( j \), for each \( j \neq i \)

The probability of having \( k \) matches (successes) out of \( N \)
SNPs (trials) follows the binomial distribution.

\[
p_k = \binom{N}{k} p^k q^{N-k} = \frac{N!}{k!(N-k)!} p^k q^{N-k} \sum_{k=0}^{N} p_k = 1
\]

where

\[
p = P_{-pm} \quad q = P_{-mm} \quad (q + p = 1)
\]
Gaussian approximation of Binomial distribution

Given \( p \) and \( N \):

\[
\begin{align*}
    k_{\text{mean}} &= Np \\
    sd_{k_{\text{mean}}} &= \sqrt{Np(1-p)}
\end{align*}
\]

Thresholds can be set as:

\( k_{\text{thres}} = k_{\text{mean}} + /- m \cdot sd_{k_{\text{mean}}} \)
PROBABILISTIC APPROACH
DOUBLE TEST

GOLD STANDARD POPULATIONS

Not pair population \( p_{pm} 0.4 \) \( m_{sig} 2 \)

Pair Population \( p_{pm} 0.9 \) \( m_{sig} 1 \)

Varying the \( m_{PAIR} \) and \( m_{NON-PAIR} \) we set how CONSERVATIVE the test is.

3/16/09 F Demichelis WCMC
Distance $D$ (% of discordant genotype calls)

- 5.3K SNPs
- 80 SNPs
- 60 SNPs
- 40 SNPs
- 20 SNPs

Index of Cell Line Pairs

OVACR-8 and NCI-ADR-RES

SPIA TEST: different, uncertain, similar

less than 90% of available calls
Distance $D$ (% of discordant genotype calls)

SPIA TEST: different, uncertain, similar

MCF7 and BT20
Primary & Met
Same patient? UNEXPECTED
Xeno & rapid autopsy tissue from the same patient

3/16/09 F Demichelis WCMC
**PROSTATE CELL LINES (from Jill Macoska)**

At DIFFERENT PASSAGES

<table>
<thead>
<tr>
<th>Name</th>
<th>#_Passage</th>
<th>#_Tubes</th>
<th>Aliquote in lab</th>
<th>50KXba</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 N15C6</td>
<td>48</td>
<td>5</td>
<td>2.3</td>
<td>60106</td>
</tr>
<tr>
<td>3 N15C6</td>
<td>50</td>
<td>7</td>
<td>17</td>
<td>60106</td>
</tr>
<tr>
<td>5 N15C6</td>
<td>52</td>
<td>3</td>
<td>2.5</td>
<td>60106</td>
</tr>
<tr>
<td>7 N15C6</td>
<td>54</td>
<td>26</td>
<td>1.7</td>
<td>60106</td>
</tr>
<tr>
<td>9 N15C6</td>
<td>56</td>
<td>9</td>
<td>1.25</td>
<td>60106</td>
</tr>
<tr>
<td>11 N15C6</td>
<td>58</td>
<td>21</td>
<td>7.7</td>
<td>60106</td>
</tr>
<tr>
<td>13 N15C6</td>
<td>60</td>
<td>19</td>
<td>1</td>
<td>60106</td>
</tr>
<tr>
<td>15 N15C6</td>
<td>63</td>
<td>24</td>
<td>0.9</td>
<td>60106</td>
</tr>
<tr>
<td>1 N33B2</td>
<td>21</td>
<td>2</td>
<td>2.14</td>
<td>60106</td>
</tr>
<tr>
<td>4 N33B2</td>
<td>27</td>
<td>4</td>
<td>2.2</td>
<td>60106</td>
</tr>
<tr>
<td>8 N33B2</td>
<td>33</td>
<td>1</td>
<td>1.25</td>
<td>60106</td>
</tr>
<tr>
<td>10 N33B2</td>
<td>35</td>
<td>11</td>
<td>2</td>
<td>60106</td>
</tr>
<tr>
<td>12 N33B2</td>
<td>37</td>
<td>20</td>
<td>19</td>
<td>60106</td>
</tr>
<tr>
<td>13 N33B2</td>
<td>39</td>
<td>14</td>
<td>1</td>
<td>60106</td>
</tr>
</tbody>
</table>
SPIA – Allelotype distance

PROSTATE CELL LINES (from Jill Macoska)
At DIFFERENT PASSAGES

50K chip – 58960 SNPs

Between N15C6 and N33B2 or Control
Between Control and passages of N33B2
Between passages of N15C6
Between passages of N33B2

3/16/09
F Demichelis WCMC
SPIA – Allelotype distance

PROSTATE CELL LINES (from Jill Macoska)

At DIFFERENT PASSAGES

SPIA top 54 selected SNPs

Passages_Macoska_SNP_data_060306 All samples (15) on 54 SNPs - Sel Top 54

3/16/09

F Demichelis WCMC
SPIA – Allelotype distance

Chromosome 11 - 2889 SNPs

PROSTATE CELL LINES (from Jill Macoska)
At DIFFERENT PASSAGES

Up to 20% difference between p48 and p63

3/16/09
F Demichelis WCMC
Our interest is in studying genome polymorphisms with respect to cancer susceptibility and characterization, by applying quantitative methods to genome-wide data.