Genotyping with SNP chips

Contributors to this lecture: Benilton Carvalho and Terry Speed

What are SNPs?

- SNPs make up 90% of all human genetic variations, and SNPs with a minor allele frequency of ≥ 1% occur every 100 to 300 bases along the human genome, on average.
- Variations in the DNA sequences of humans can affect how humans develop diseases, respond to pathogens, chemicals, drugs, etc. As a consequence SNPs are of great value to biomedical research and in developing pharmacy products.

From Wikipedia

Remember

- You have two alleles: From mom and from dad
- Each one is either A or B, so you can be AA, AB, BB
- Our task is to use microarrays to know genotype for 1000s SNPs at a time
- Remember: DNA has to strands: sense (+) and antisense (-)
**The Affymetrix genotyping microarray**

Whole Genome Sampling Assay

1. **Fractionate** total genomic DNA with a restriction enzyme (e.g. XbaI)
2. **Ligate** a single generic adaptor to the ends of all fragments
3. Use the generic adaptor as primer pair to carry out the PCR, amplifying fragment sizes (250 bp - 2,000 bp) such that the
   - PCR is reliable and reproducible, and the
   - Total PCR product is small enough to hybridize efficiently
4. Fragment, label, hybridize, stain, wash, scan, analyze image, then analyze data to call genotypes (our task).

**Single primer assay: complexity reduction**

**Single primer assay: overview**
Affymetrix SNP chip terminology

Genomic DNA
TAGCCATCGGTAGTACTCAATGAT

Perfect Match probe for Allele A
ATCGGTAGCCATCATGAGTTACTA

Perfect Match probe for Allele B
ATCGGTAGCCATCATGAGTTACTA

Genotyping: answering the question about the two copies of the chromosome on which the SNP is located:
Is a person AA, AG or GG at this Single Nucleotide Polymorphism?

Tiling strategy

SNP position 0
A / G
TAGCCATCGGTA N GTACTCAATGAT

| PM 0 Allele A | ATCGGTAGCCAT | T | CATGAGTTACTA |
| MM 0 Allele A | ATCGGTAGCCAT | A | CATGAGTTACTA |
| PM 0 Allele B | ATCGGTAGCCAT | C | CATGAGTTACTA |
| MM 0 Allele B | ATCGGTAGCCAT | G | CATGAGTTACTA |

Central probe quartet

Tiling strategy, 2

SNP Position +4
A / G
TAGCCATCGGTA N GTA C TCAATGATCAGCT

| PM +4 Allele A | GTAGCCAT | T | CAT | G | AGTTACTAGT CG |
| MM +4 Allele A | GTAGCCAT | T | CAT | C | AGTTACTAGT CG |
| PM +4 Allele B | GTAGCCAT | C | CAT | G | AGTTACTAGT CG |
| MM +4 Allele B | GTAGCCAT | C | CAT | C | AGTTACTAGT CG |

+4 offset probe quartet
In summary: probe level data

- Two alleles
- Two directions
- Two types (PM,MM)
- Up to 7 locations of the SNP in the probe

Affymetrix SNP probe tiling strategy, 3

<table>
<thead>
<tr>
<th>Offset quartets</th>
<th>Central quartet</th>
<th>Offset quartets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PMA</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>MMAPMM</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>MMAM</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>MMAM</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>MMAM</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>MMAM</td>
<td>7</td>
</tr>
</tbody>
</table>


Probe Intensities

Fake (idealized) image for 3 samples on one SNP

Fake, as the probes are not all adjacent on the chip idealized, as all the probes are high or low as they should be.
Calling genotypes:  
A modular approach

**MPAM:** The first Affymetrix SNP-calling algorithm, used on the 10K SNP chip

Generalities concerning MPAM

- Derive a reasonable though ad hoc summary statistic, here RAS (feature extraction)
- Clusters the statistic in a sensible way, here using MPAM (classification)
- Generates new calls by cluster membership, here using elliptical regions, cf. bivariate normal (modeling).

Ref: Liu, WM et al, *Bioinformatics* Dec 2003

MPAM: detection filter

\[
\begin{align*}
  i & \in \{S,T\} \quad \text{Sense or antisense strand} \\
  j & \in \{A,B\} \quad \text{allele} \\
  k & \in \{1,\ldots,7\} \quad \text{position of interrogation} \\

  D_{ik} & = \frac{(PM_{ik} - MM_{ik})}{(PM_{ik} + MM_{ik})} \\
  D & = \text{median}(D_{ik}) \\
  D & = \max(\min(D_{1A},D_{1B}), \min(D_{2A},D_{2B})) \\

  \text{SNPs with low } D (<0.03) \text{ are not called.}
\end{align*}
\]
**MPAM: feature extraction**

- **i** ∈ {S, T} Sense or anTisense strand (also +, - or 1,2)
- **j** ∈ {A,B} allele
- **k** ∈ {1, ..., 7} position of interrogation

\[
MM_k = (MM_{ik} + MM_{jk})/2 \\
S_k = \text{max}(PM_{ik} - MM_k, 0) \\
s_{ik} = \text{Relative Allele Signal of k-th quartet of strand i} \\
s_{ik} = \frac{s_{ik, A}}{s_{ik, A} + s_{ik, B}} \\
s_i = \text{Relative Allele Signal of strand i} \\
s_i = \text{median}(s_{ik})
\]

**Clustering and modeling**

- **MPAM: classification algorithm**

  - **Partitioning Around Medoids PAM**
    - Kaufman and Rousseeuw, 1987
  - **Work with Relative Allele Signal RAS (s_A, s_T), 2-dim feature space from both forward and reverse strands**
  - **n** points in feature space: \(x_1, x_2, ..., x_n\)
    - Assuming there are \(k = 2\) and \(3\) groups
      - minimize \(f_{PAM} = \sum_{i=1}^{n} (\text{min}(d(x_i, m_j), j=1:k))\)
  - **MPAM (modified PAM): minimize**
    - \(f_{MPAM} = f_{PAM} - \sum_{i=1}^{n} (\text{min}(d(x_i, x_j), x_j \in G_p, x_i \notin G_p))\)
**Difference between PAM and MPAM**

The result of using PAM with 3 groups on the data for one SNP.

The penalty used on MPAM is designed to avoid just this situation.

---

**Genotyping using robust models**

---

**MPAM Classification quality metrics**

Silhouette width for $x_i$

- $a(i)$ is the average within-group distance to $x_i$
- $b(i)$ is the average between-group distance to $x_i$
- $w(i) = (b(i) - a(i))/\max(b(i), a(i))$
- $w = \text{average}(w(i), i=1...n)$

Separation of the groups

- $sep_{\text{median}}(|x_{aa} - x_{ab}|, |x_{ab} - x_{bb}|)$
- $sep_{\text{median}}(|y_{aa} - y_{ab}|, |y_{ab} - y_{bb}|)$
- $sep = \min(sep_{\text{median}}, sep_{\text{median}})$
Worked fine for the 10K

- 99.5% accuracy
- 99.998% reproducibility
- 97% call rate

Why not MPAM for 100K?

- Large sample size is needed for clustering
- Hard to handle SNPs with low minor allele frequency: estimating location for missing genotypes is difficult.
- Visual inspection is impossible
- Models are empirical, hard to make further improvements after product launch - any changes including experimental conditions, scanner settings etc., will force rerun of experiments and rebuilding of models

Gentle critique of MPAM

- RAS ad hoc...why this rather than another measure? (Possible answer: it works!)
- The procedure makes no use of many features of the data, most importantly the known genotypes, and repeatable probe behaviour
- Fails to exploit the massive parallelism inherent in the 100K SNP chip.
Unified approach: the Dynamic model-based algorithm, DM

Until recently the vendor-supplied genotype-calling algorithm. Seeks the best fitting pattern of the above kind, including no call (NC). It is a mix of normal likelihood-based model selection and a Wilcoxon test, leading to a final p-value which is a form of confidence statement about the call.

There is no training, and it is a single chip procedure.

However, the SNPs on the chip have been selected so that the algorithm works on them.

DM

• Look at quartets individually and produce a score under normal theory assumptions
• Combine scores across quartets to produce a classification into genotypes (resistant to cross-hybridization and model failure)
• Provides a “p-value”/goodness of classification metric

Ref: Di, X. et al, Bioinformatics May 2005

Likelihood, intensity scale, for each quartet

$$\prod_{i \in \{A,C,G,T\}} n_i \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left\{ -\frac{1}{2} \left( \frac{x_{i,j} - \mu_i}{\sigma_i} \right)^2 \right\}$$

$$n_i = \text{number of pixels for feature } i; \ x_{i,j} = \text{measured value of pixel } j; \ \mu_i = \text{unknown mean pixel intensity}; \ \sigma_i = \text{unknown SD of pixel intensities, all for feature } i \in \{A,C,G,T\}; \ x' \text{ denotes reverse strand.}$$

Null model (B for background)

$$\mu_A = \mu_C = \mu_G = \mu_T = \mu_B; \ \mu'_A = \mu'_C = \mu'_G = \mu'_T = \mu'_B$$

Illustrative homozygote model: CC (S for signal)

$$\mu_C = \mu'_B; \ \mu_A = \mu_G = \mu_T = \mu_B; \ \mu'_C = \mu'_A = \mu'_G = \mu'_T = \mu'_B$$

Illustrative heterozygote model: CT

$$\mu_C = \mu_T; \ \mu_A = \mu_G = \mu_B; \ \mu'_C = \mu'_T = \mu'_A = \mu'_G = \mu'_B$$
**DM: combining quartet-level information**

- Start with \( N \) probe quartets \( q_i \), \( i = 1, \ldots, N \), \( N \) typically 10 or 14
- For each probe quartet \( q_i \), evaluate log-likelihood \( LL \) of the 4 possible models:
  - \( LL_{AA,i} \), \( LL_{AB,i} \), \( LL_{BB,i} \), \( LL_{NC,i} \), \( NC = \text{No Call} \)
- For each probe quartet, transform log-likelihoods to scores:
  - \( s_{AA,i} = LL_{AA,i} - \max\{LL_{m,i} \mid m \neq AA\} \)
  - \( s_{AB,i}, s_{BB,i}, s_{NC,i} \) computed similarly
- Combine quartet-level results to a SNP-level result:
  - for each model \( m \in \{AA, AB, BB, NC\} \) use Wilcoxon signed rank test on \( \{s(m,i) ; i = 1, \ldots, N\} \)
  - Yields 4 \( p \)-values, the call and score for the SNP corresponds to the model with the most significant \( p \)-value

---

**DM on 30 CEPH trios: HapMap Concordance & Mendelian Inheritance**

---

**Why attempt an improvement over DM?**

- Perhaps the error rate is too high?
- There is reason to believe it can be improved by
  - a) using the training/test set paradigm;
  - b) carrying out multi-chip analyses, which identify and exploit probe behaviour; and
  - c) exploiting the massive parallelism across SNPs.
- The 100K SNPs were selected from a much larger screening set using DM. For the 500K and >1M SNP chips, a higher yield is desirable, and perhaps a better genotype-calling algorithm could achieve this.
Robust Linear Model with the Mahalanobis distance classifier

- RLMM pronounced “REALM”
- Based on an RMA-like model
  - Uses PM only
  - Linear additive multi-chip model on log scale
  - A- and B-probe and chip effects
  - Robustly estimated parameters
- Classification using Mahalanobis’ distance
- Morphed into BRLMM; CRLMM coming up!

Notation

- Once we are done with first part of preprocessing we have the following:
  \[ \theta_A \text{ and } \theta_B \text{ proportional to log of the amount of fragments from allele A and B respectively} \]
  
  In principal these can only be (log of) 0, x, or 2x, but we know better than to believe this. In fact we know not to expect the same cut-off to work for all SNPs

It’s not easy

This picture shows that most the information is in the left right diagonal direction, i.e. in the log-ratios
Why is this?

- Our guess is that the PCR step introduces a lot of SNP to SNP variation
- We have proxies for measuring PCR effect: fragment sequence and fragment length
- We can examine the fragment sequence via the probe sequence

Sequence effect
Normalization

• We normalize/summarize using RMA (no BG correction) after correcting for sequence and length effects on the log intensities
• We then examine log-ratios
• We keep sense and antisense separate

“Broken” probes (BRLMM)

“Broken” probes?
Log-ratio biases persist

Different arrays, different cut-offs

Length effect on M
Intensity effect on M

Use mixture model to fix this

\[ M_i | Z_i = k = f_k(X_i) + \varepsilon_{i,k} \]

- SNP denoted with \( I \)
- \( Z \) is true, so \( k = AA, AB \) or \( BB \)
- \( X \) are covariates that cause bias

After fix
After our normalization

General Improved Separation

Fragment length effect
“Broken” probes (RLMM)

Preprocessing model motivates genotype algorithm

\[ [M_{k,j}] = k, m_{j,k} = f_{j,k}(X_{k,j}^2) + \epsilon_{j,k} \]

- Array denoted with $j$
- Shift in cluster center denoted with $m$
- Assume $m$ are bivariate normal and
  \[ \frac{1}{\alpha_{ij}} = \frac{1}{\alpha_{ij}} \cdot \epsilon_{ij} \]
- Use training data to estimate
- Use empirical bayes approach for cases with few data points

Predicting regions

\[ \begin{align*}
\hat{m}_k &= (\nu^{-1} + N_k \Sigma^{-1})^{-1} N_k \Sigma^{-1} \hat{m}_k \\
\hat{\Sigma}_{k,k}^2 &= \frac{(N_k - 1) \hat{\sigma}^2_{k,k} + d_{k} \hat{\sigma}^2_{d,k}}{(N_k - 1) + d_{k}}, \text{ for } N_k > 1.
\end{align*} \]
Example