Ambiguous Probes Create Artificially High Correlation of Expression levels between Affymetrix Probe-sets

Renaldo Webb

Summer Undergraduate Research Program in Genomics

Broad Institute

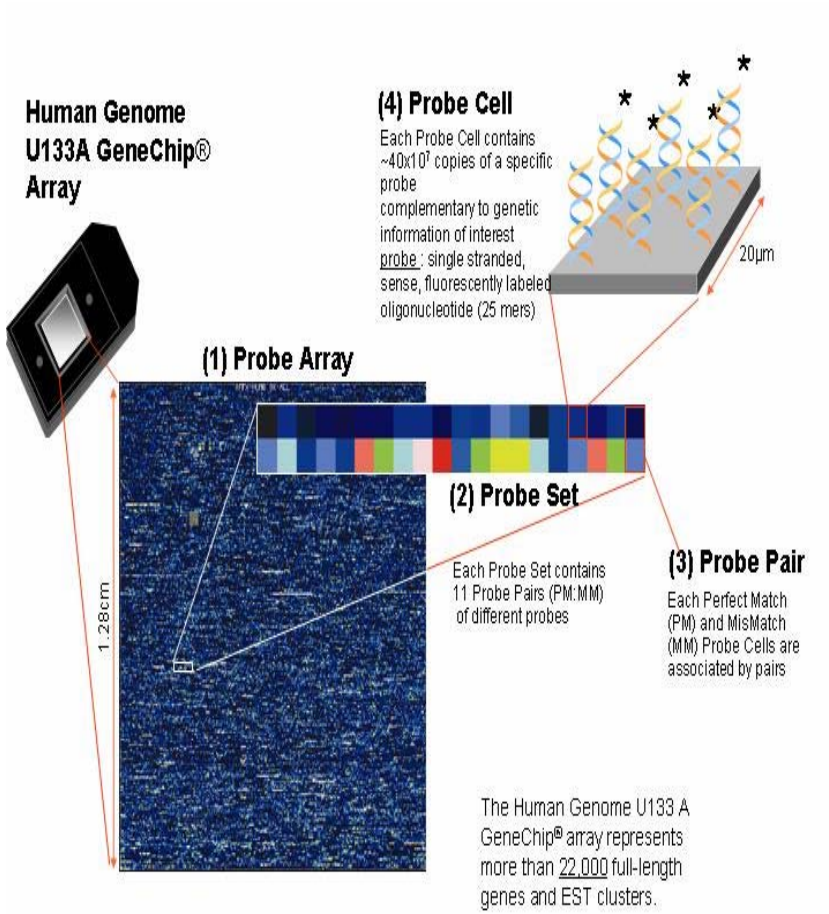
Affymetrix GeneChip

- Powerful genome-wide expression tool
- First DNA microarray able to analyze the entire genome
- Understanding disease

http://upload.wikimedia.org/wikipedia/commons/2/22/Affymetrix-microarray.jpg

How Does It Work?

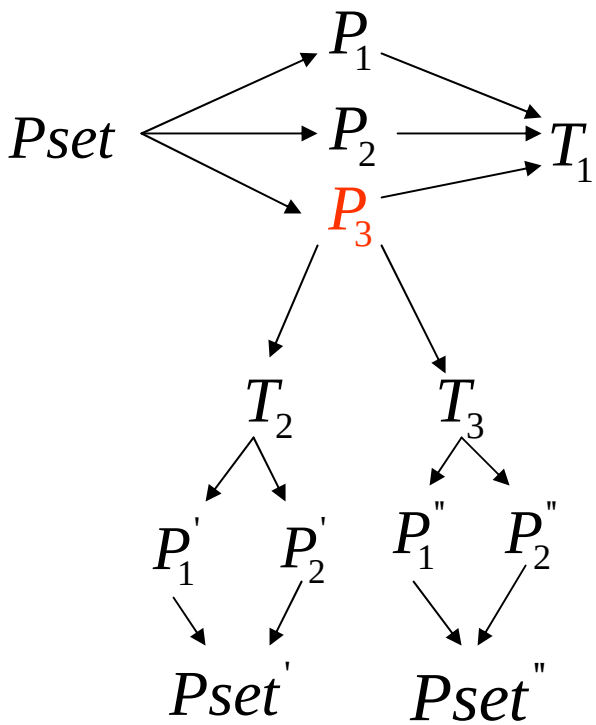
- Probe-sets contain 11- 20 probes
- Probes are 25 base pairs long
- Probes bind to target transcripts
- What happens if probes bind to more than one transcript?



http://www.weizmann.ac.il/home/ligivol/research_interests.html

Ambiguous Probes

- Probes should be unique
- Ambiguous probes bind to multiple transcripts
- Affects expression data
- Genome-wide sequencing mapping



Correlation

- Relationship between two independent variables
- Natural correlation
- Ambiguous probes would create an artificially high correlation
- Can we find such signatures?

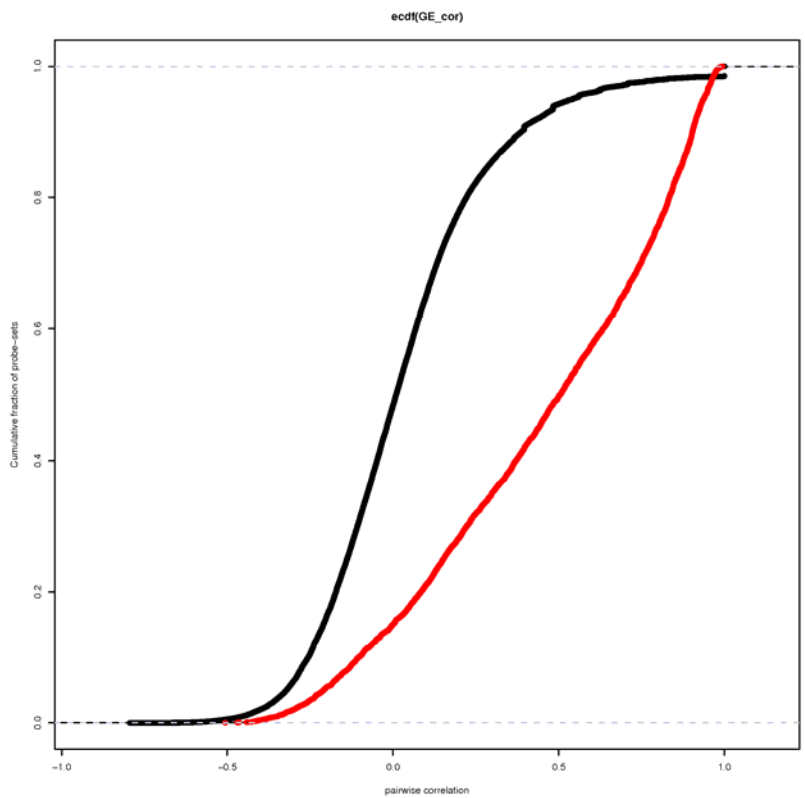
Finding Ambiguous Probes

- Original Algorithm in Perl:
  - Read in Affymetrix Probe-set definitions
  - Organize data into 4 main hashes
  - Use if statements to find probes with multiple transcripts
  - Stored probes in an array and use in conjunction with hashes to find probe-sets and transcripts
  - One program organized this data into CDF file
  - The second organized this data into a vector

Finding Correlations

- Algorithm in R:
  - Read in vector created by Perl and breast cancer expression data
  - Organize the vector of probe-sets into an nxn matrix
  - Calculate baseline correlation
  - Calculate actual correlation
  - Use CDF file to remove ambiguous probesets
  - Calculate the new correlation
  - Graph and save results

Initial Results



- Black- baseline correlation

$$Pset \longrightarrow T_1$$

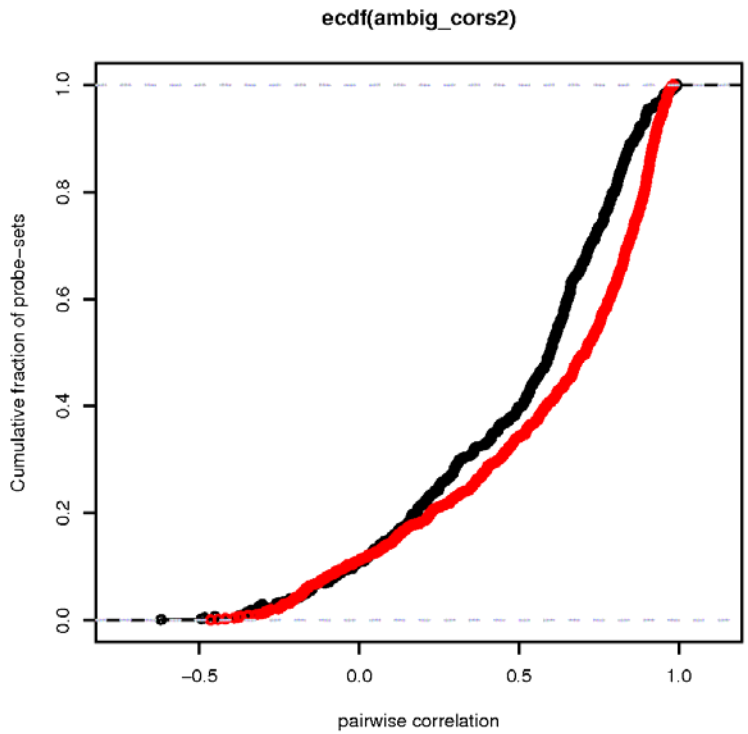
$$Pset' \longrightarrow T_2$$

- Red- correlation with ambiguous probes

$$Pset \rightarrow \begin{cases} T_1 \\ T_2 \end{cases}$$

$$Pset' \longrightarrow T_2$$

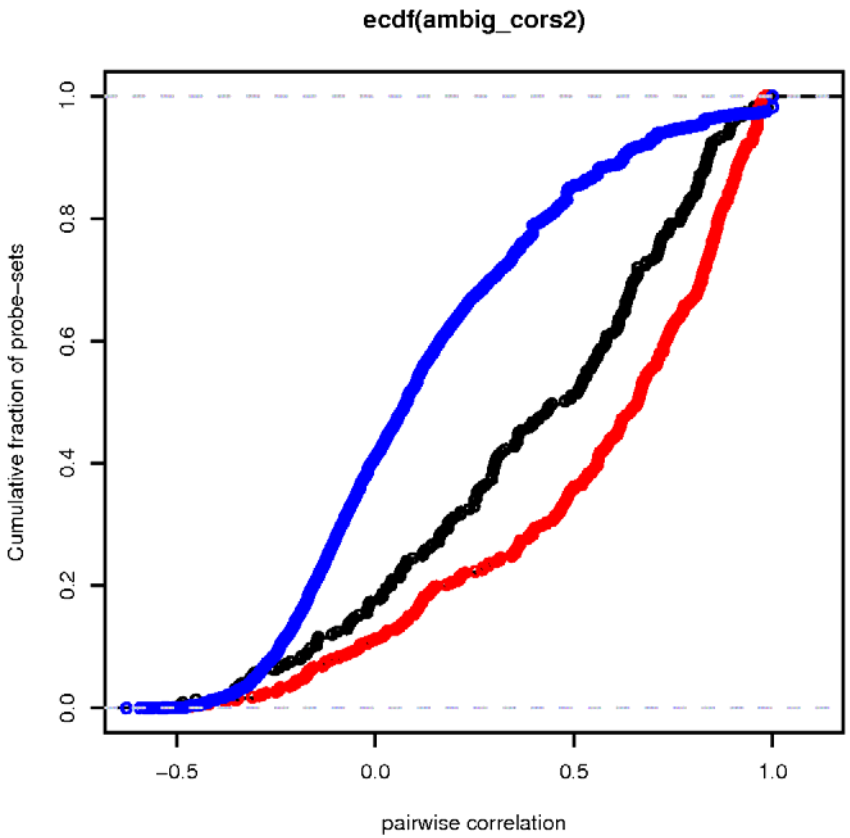
Once Ambiguous Probes are Removed



- Black- no ambiguous probes
- Red- ambiguous probes
- The CDF file did not remove:

$$Pset \rightarrow P_1 \rightarrow T_1$$
$$Pset \rightarrow P_2 \rightarrow T_2$$

Final Results



- Blue- baseline
- Black- no ambiguous probes
- Red- ambiguous probes
- Probe-sets that measured 8 or more transcripts were removed

Other Interesting Findings

- The HGU133A chip has 12,917 ambiguous probes- 10%
- 2,150 of probe-sets in the HGU133A chip have no unique probes- 9%
- 128 probe-sets target eight or more transcripts- 1%

Conclusion

- Ambiguous Probes cause artificially high correlation in expression levels
- Correlation is a function of ambiguity
- Removing ambiguous probes will decrease systematic experimental error.

Acknowledgements

- Scott Carter
- Angela Brunache
- Bruce Birren
- Eric Lander
- The Broad Institute