Genome Scale Family Based Association Testing using Condor

Alex Stoddard
astoddard@mcw.edu
MCW Bioinformatics Program
April 21st, Condor Week 2009
“Something old, something new ...” (Traditional English rhyme)

- Much theory for genetic mapping predates the discovery of DNA
- Within the last decade the human genome and haplotype map have become available.
“We have the technology ...” (The Six Million Dollar Man)

With DNA microarrays we can assay a million DNA variants (SNPs – single nucleotide polymorphisms) per individual.

Still a factor of a thousand smaller than the actual genome sequence.

We can expect within another 10 years to have affordable individual genome sequences.

The human genome project cost about $3 billion.

Your doctor might soon be able to order your personal genome for $300.

But will he know what it means?
Simple vs. Complex Traits (phenotypes, diseases)

Simple, or Mendelian Traits,

- e.g. Cystic Fibrosis
- Sickle-cell anemia
- Haemophila A & B (two different genes factors VIII/IX)

Complex Traits (non-Mendelian)

- Type II Diabetes
- Hypertension
- Cardiovascular diseases

Cumulative effect of many genes potentially interacting with multiple environmental factors.

"Make everything as simple as possible, but not simpler."

(Albert Einstein)
Keep it in the family...

- **HyperGEN study** (National Heart, Lung and Blood Institute NHLBI).

Over 100 cardiovascular and blood pressure related traits assayed.
“Even the strongest man cannot lift a heavy heart.”

Even the strongest man cannot lift a heavy heart.

(Chinese proverb)

Left Ventricular Hypertrophy (LVH)

A heritable, independent risk factor for cardiovascular death.
“Scotty, We Need More Power!”  
(Captain Kirk *Star Trek*)

- Genome scale analyses are mostly done without using the available family structure.
  - Computationally efficient but requires extra assumptions and model corrections.

- Multiple family based analysis algorithms exist with different implementations (PBAT, QTDT etc.).
  - Analysis is very parallelizable, but software was written before genome scale data was available.
  - Testing one simple set of model parameters for 500,000 SNPs in hundreds of families will take about 40,000 cpu/hrs.
  - Condor is a great fit.
“In nature’s infinite book of secrecy. A little I can read.”
(Shakespeare, Antony and Cleopatra)

• Privacy concerns when working with human genome data.

An individual’s genome is inherently both biomedical and biometric.

De-identification cannot work here.
“A journey of a thousand miles begins with a single step.”
(Lao-tzu, Chinese philosopher)

• Ongoing development:
  – Management of batch analysis
  – Prioritizing SNPs for testing multiple models and analysis methods (DAGMan?)
  – Building condor pool at MCW
  – “Flocking” to other area resources.
Acknowledgments

Medical College of Wisconsin
Ulrich Broeckel
Rachel Lorier
Andrea Matter
Beth Virlee
Karen Maresso
George Kowalski
Greg McQuestion
... and many others in IMI, HMGC, & BP groups

University of Alabama at Birmingham
Donna Arnett
Hemant Tiwari
Amit Patki

Washington University
DC Rao

University of Wisconsin, Madison
Miron Livny
Condor Group