FaST Algorithms for Extremely Large Genome-Wide Association Studies

David Heckerman

Christoph Lippert, Jennifer Listgarten,
Bob Davidson, Carl Kadie

Microsoft Research

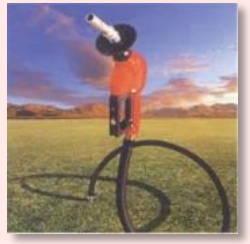
eScience Research Group in MSR

Tackling societal challenges with machine learning

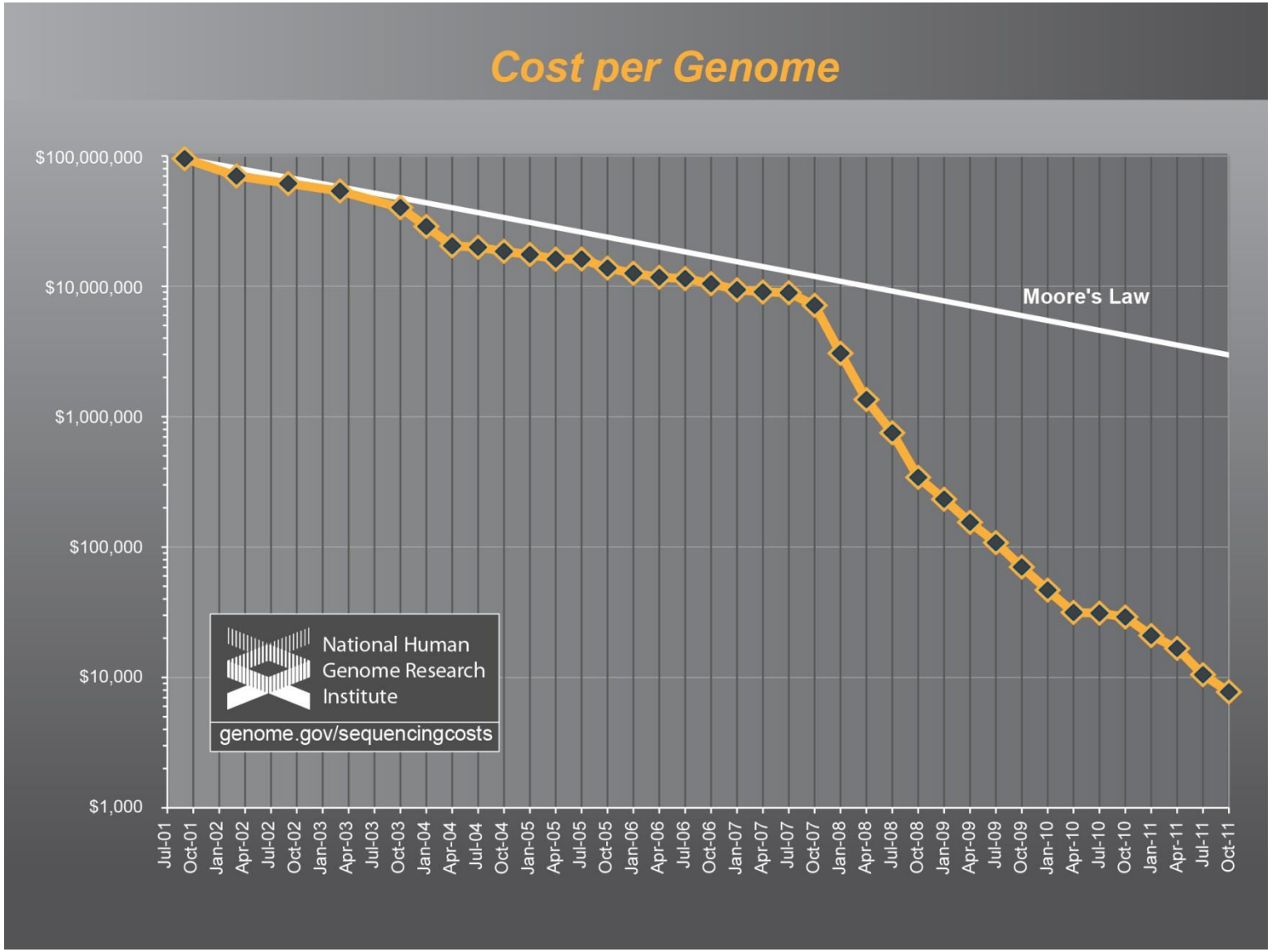
Human genomics and personalized medicine

Vaccine design
HIV, HCV, HRV

Genome assembly of sugar cane



The Genomics Revolution



Personalized Medicine

Use genetic markers to…

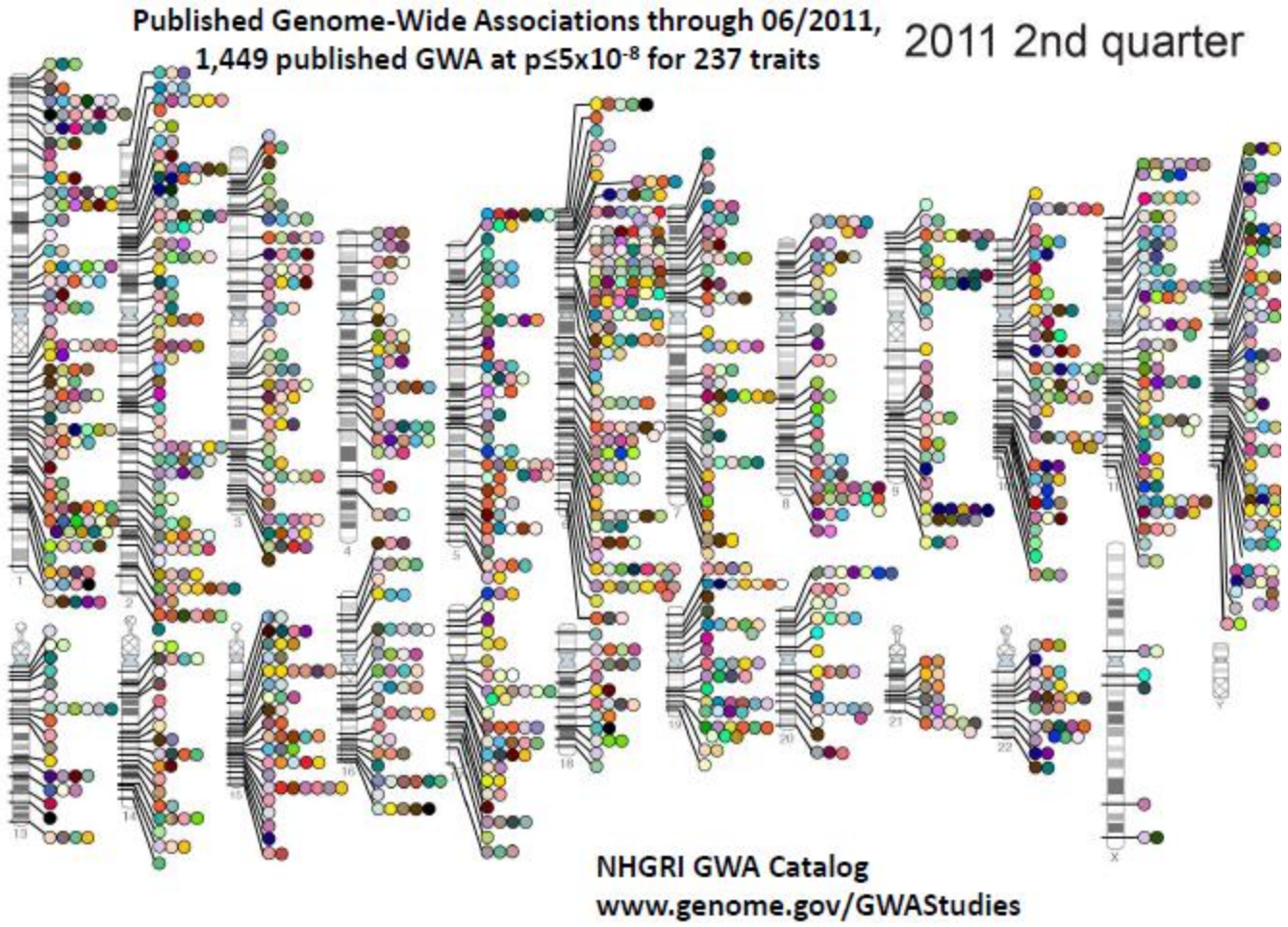
- Understand causes of disease
- Diagnose a disease
- Infer propensity to get a disease
- Predict favorable response to a drug
- Predict bad reaction to a drug

GWAS: Genome-wide association studies

...actgccgcgaC actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaG actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaC actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaC actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaC actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaC actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...

...actgccgcgaG actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaG actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaC actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaG actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaG actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...
...actgccgcgaG actgccgcgaggcctgactggcatccagtttagcggaactgccgcaa...

Genome-Wide Association Studies

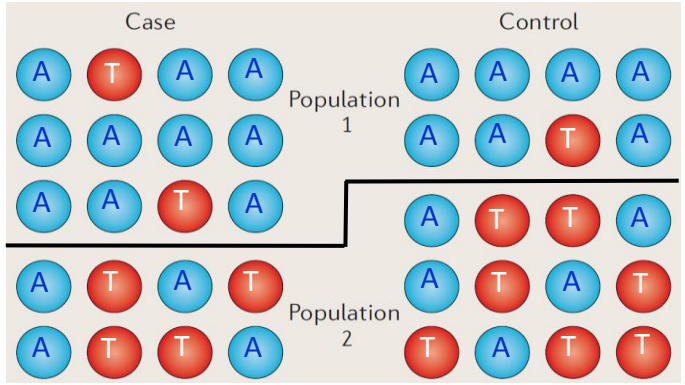


GWAS issues

- For common diseases, signals are spread out across the genome and weak
- To pick up these signals, we need lots of data
  - deCode
  - 23andMe
  - Kaiser
- Large data → confounding
  - Multiple ethnicities
  - Related individuals

Challenge: Confounding factors

- Suppose the set of cases has a different proportion of ethnicity X from control.
- Suppose we use linear regression to look for SNP-trait correlations.
- Then genetic markers that differ between X and other ethnicities in the study, Y, will appear artificially to be associated with disease.
- Problem gets worse with more data.



One solution: Throw out data

- Use one ethnicity
- Discard data for individuals who are too closely related

Bad idea: The most powerful comparisons are between closely-related individuals

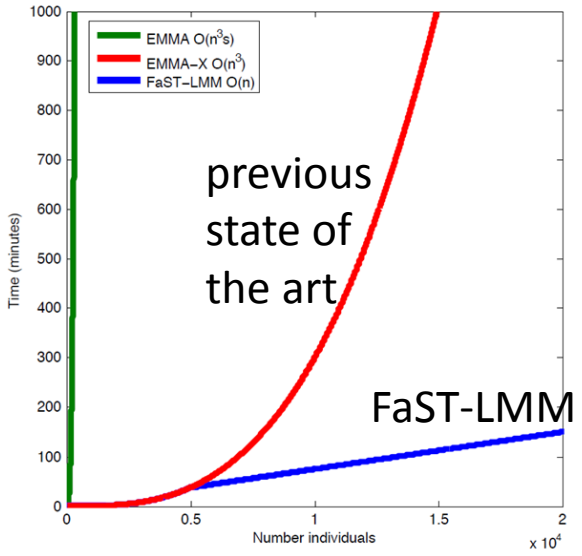
FaST-LMM: Factored Spectrally Transformed Linear Mixed Models

Linear mixed models:

- $O(N^3)$ runtime
  $O(N^2)$ memory use
  N<5,000
- FaST-LMM has O(N) runtime and memory use; N>100,000
- More power than traditional approach

nature | **methods** September 2011

nature | **methods** June 2012



Faster and more power: Really?

- LMM captures confounding structure by computing similarities between all pairs of individuals
- In this view, one would think all SNPs should be used
- BUT: LMM is equivalent to performing linear regression with these SNPS as covariates
- If someone gave you a million potential covariates, would you blindly use them all?
- Solution: Variable selection (“feature selection”)
- Result: ~100s of SNPs selected for similarity computation → less SNPs than individuals → O(N)

Application: Epistasis GWAS
(SNP-SNP interactions)

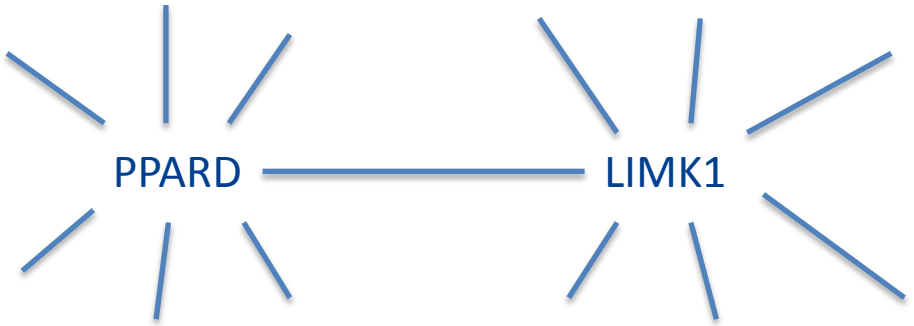
Seven common diseases:

- Type I diabetes
- Type II diabetes
- Coronary artery disease
- Hypertension
- Crohn's disease
- Rheumatoid arthritis
- Bipolar disorder

With FaST-LMM and Azure, can look at all SNP pairs (about 60 billion of them)

1,000 compute years; 20 TB output – we did it in 13 days on Azure

New results for CAD



SCIENTIFIC REPORTS

An Exhaustive Epistatic SNP Association Analysis on Expanded Wellcome Trust Data

Christoph Lippert, Jennifer Listgarten, Robert I. Davidson, Jeff Baxter, Hoifung Poon, Carl M. Kadie & David Heckerman

Software freely available for academic use (just bing “FaST-LMM”)