Challenges and requirements for Biostatistics/Bioinformatics applications

Benno Pütz
Lifetime prevalence

~20%
Treatment response

Binder et al., Nature Genetics 2004
Drug transport

Dosierung

MPI of Psychiatry
Drug transport

Substrate des P-Glykoproteins
Amitriptylin, Citalopram, Paroxetin, Venlafaxin

Zeit [Wochen]

C Träger (N=23)
Nicht-C-Träger (N=110)

Uhr et al., Neuron, 2008
150 years ago
Inheritance
Heritability
Hypothesis-driven research
candidate genes
Problem
Complex Diseases
many genes
weak contributions
heritability rate
Genetic contribution

<table>
<thead>
<tr>
<th>Category Title</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipolar Disorder</td>
<td>85%</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>80%</td>
</tr>
<tr>
<td>Unipolar Depression</td>
<td>40%</td>
</tr>
</tbody>
</table>

(MPI of Psychiatry)
environment
observation
treatment response
identify mechanisms
understand processes
genetic level
targets for intervention
individually tailored
Personalized Medicine
unbiased approaches
Gene → Transcript → Protein → Trait
Genome \rightarrow Transcriptome \rightarrow Protein \rightarrow Trait
Genome ➔ Transcriptome ➔ Proteome ➔ Trait
Genome ➔ Transcriptome ➔ Proteome ➔ Trait
High-Throughput

Genotyping

Gene Expression

Proteomics
hypotheses
unbiased approaches
High-Throughput

Genotyping

Gene Expression

Proteomics
High-Throughput

Genotyping

Gene Expression

Proteomics

NGS

MPI of Psychiatry
High-Throughput

Genotyping

Gene Expression

Proteomics

MPI of Psychiatry
Clinical limitations

- Patient data
- Privacy concerns (Bay. Krankenhausgesetz)

⇒ in-house solution
Hardware

- cluster
  - 192 cores (4 × 48)
  - 256GB RAM each
- 48 TB
Hardware

- Cluster
- GPU

GTX 295
Hardware

- Cluster
- GPU
Hardware

- cluster
- GPU
Genotyping
Genetic variations

Human genome

> 99% conserved across individuals
Genetic variations

• SNPs
Genetic variations

- SNPs
- CNVs

Graph showing the number of copies of a genetic region: 2 copies and 0 copies.
Genetic variations

- SNPs
- CNVs
- Mutations
Genetic variations

- SNPs
- CNVs
- Mutations
Genetic variations

- SNPs
- CNVs
- Mutations
Genetic variations

- SNPs
- CNVs
- Mutations
- Epigenetic

- methylated cytosine
- acetylated lysine
Challenges
Imputation
SNP Chips

Genome-wide genotyping

- Affymetrix: 10k, 100k, 500k, 500k +420k, 900k +950k
- Illumina: 100k, 300k, 500k, 1M

only partial overlap
Large studies

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-8}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn’s disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point $P$ values between $10^{-8}$ and $5 \times 10^{-7}$) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes observed at most loci identified. This study thus represents a thorough validation of the GWA approach. It has also demonstrated that careful use of a shared control group represents a safe and effective approach to GWA analyses of multiple disease phenotypes; has generated a genome-wide genotype database for future studies of common diseases in the British population; and shown that, provided individuals with non-European ancestry are excluded, the extent of population stratification in the British population is generally modest. Our findings offer new avenues for exploring the pathophysiology of these important disorders. We anticipate that our data, results and software, which will be widely available to other investigators, will provide a powerful resource for human genetics research.
SNP Chips

- Affymetrix
  - 10k
  - 100k
  - 500k
  - 2 * 250k
  - 500k
  - 900k
  - +420k
  - +950k

- Illumina
  - 100k
  - 300k
  - 600k
  - 1M

unify
Imputation

- Combine studies
  - different SNP coverage
- Extent to reference samples
  - HAPMAP project
  - 1000 genomes
    - 1,386,077
    - 7,611,725
    - 7,835,704
Observation vs. Reference

Observed Genotypes

A
G

Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A A G T T T T T C C T T T T A C
C G A A G T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
Find matches

Observed Genotypes


Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T T C T T C T G T G C
C G A A G C T C T T T T T C T T C T G T G C
C G A A G C T C T T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T C T T G T G C
C G A G A C T C T T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T T C T T C T G T G C
Phase and impute

Observed Genotypes

c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t C A t g g

Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C C
C G A A G C T C T T T T C T T C T G T G C C
C G A A G C T C T T T T C T T C T G T G C C
C G A A G C T C T C T C G A A C C T T A T G C
T G G G G A T C T C C C G A A C C T C A T G G
C G A G A T C T C C C G A A C C T T T G T G C C
C G A G A T C T C C C G A A C C T T T G T A C
C G A A G C T C T C C C G A A C C T C G T G C C
C G A A G C T C T T T T C T T T T G T G C C

MPI of Psychiatry
Example

Software

- Impute
- Mach
- Beagle
- Plink
- snpMatrix
MARS sample

1600 subjects

Illumina®

675 kSNPs

100k / 330k / 600k

MPI of Psychiatry
MARS sample

1600 subjects

7.85k SNPs

~2 weeks

48 cores

MPI of Psychiatry
rare variants
Family trees
Restless Leg Syndrome
Restless Leg Syndrome

exact solution: series of conditional likelihoods
Prediction

• (differential) diagnosis
• treatment
• treatment outcome
Variables

high-dimensional problem

find relevant subset
Decision trees
NGS
Sequence alignment
Idea

Shendure and Ji, Nat Biotech, 2008
Idea

Shendure and Ji, Nat Biotech, 2008
Idea

DNA fragmentation → In vitro adaptor ligation

Cyclic array sequencing
(>10^6 reads/array)

Cycle 1
AGA

Cycle 2
GAG

Cycle 3
TCG

What is base 1? What is base 2? What is base 3?

Shendure and Ji, Nat Biotech, 2008
Resequencing
Aligned reads
Divide and Conquer

Naturally parallel
Divide and Conquer

bwa multi-core
Searching for SNPs with cloud computing
Ben Langmead*, Michael C Schatz†, Jimmy Lin‡, Mihai Pop§ and Steven L Salzberg¶

Addresses: *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA. †Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA. §The iSchool, College of Information Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: blangmea@jhsph.edu

Published: 20 November 2009

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/11/R134

© 2009 Langmead et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract
As DNA sequencing outpaces improvements in computer speed, there is a critical need to accelerate tasks like alignment and SNP calling. Crossbow is a cloud-computing software tool that combines the aligner Bowtie and the SNP caller SOAPsnp. Executing in parallel using Hadoop, Crossbow analyzes data comprising 38-fold coverage of the human genome in three hours using a 320-CPU cluster rented from a cloud computing service for about $85. Crossbow is available from http://bowtie-bio.sourceforge.net/crossbow.
De-novo sequencing

- no reference sequence to align to
- computationally more demanding
Epistasis
Epistasis
Model

Linear regression

Phenotype $\sim \beta_0 + \beta_1 \text{SNP}_1$
Approximation

correlation between two SNPs

Correlation coefficients

cases

controls

Implementation in \texttt{R} using \texttt{gputools}

\texttt{cor(x, y, \ldots)}
Validation

Explicit test for "good" interactions

explicit logistic regression
Application

Multiple Sclerosis
FastEpistasis versus correlation coefficient difference p-values

- Threshold 8.680e-10

Overestimating
Epistasis: GPU

- Very fast approximation
- Very fast correlation for continuous phenotype
- Applicable to eQTL studies
→ Strong dependence on problem
→ Serious efforts required
Problems

Classification
SNP Imputation
Higher-order interactions
Epistasis
Family trees
Sequence alignment
eQTL
Problems

- Multi-core
- Sequence alignment
- SNP Imputation
- eQTL
- GPU-Parallel
- Classification
- Epistasis
- Too big/complex?
- Higher-order interactions
- Family trees

MPI of Psychiatry
Thanks
Thank you