20th International Workshop on Methodology of Twin and Family Studies: Advanced course

- Lon Cardon (director)
- Jonathan Flint
- Jeff Barrett
- David Evans
- William Valdar
- Goncalo Abecasis
- Mike Neale
- Hermine Maes
- Sarah Medland
- Dorret Boomsma
- Danielle Posthuma
- Meike Bartels
- John Hewitt (host)
- Jeff Lessem
- Matt McQueen
- Pak Sham
- Stacey Cherny
- Ben Neale
- Shaun Purcell
- Manuel Ferreira
- Nick Martin
- Kate Morley
- Marleen de Moor
- Lannie Ligthart
Hunting QTLs

Nick Martin
Queensland Institute of Medical Research

Boulder workshop: March 5, 2007
<table>
<thead>
<tr>
<th>Year</th>
<th>Location</th>
<th>Type</th>
<th>#Faculty</th>
<th># Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC1</td>
<td>1987</td>
<td>Leuven</td>
<td>Introductory</td>
<td>10</td>
</tr>
<tr>
<td>TC2</td>
<td>1989</td>
<td>Leuven</td>
<td>Introductory</td>
<td>11</td>
</tr>
<tr>
<td>TC3</td>
<td>1990</td>
<td>Boulder</td>
<td>Introductory</td>
<td>11</td>
</tr>
<tr>
<td>TC4</td>
<td>1991</td>
<td>Leuven</td>
<td>Introductory</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Advanced</td>
<td>12</td>
</tr>
<tr>
<td>TC5</td>
<td>1993</td>
<td>Boulder</td>
<td>Introductory</td>
<td>13</td>
</tr>
<tr>
<td>TC6</td>
<td>1994</td>
<td>Boulder</td>
<td>Introductory</td>
<td>16</td>
</tr>
<tr>
<td>TC7</td>
<td>1995</td>
<td>Helsinki</td>
<td>Introductory</td>
<td>10</td>
</tr>
<tr>
<td>TC8</td>
<td>1996</td>
<td>Boulder</td>
<td>Introductory</td>
<td>10</td>
</tr>
<tr>
<td>TC9</td>
<td>1997</td>
<td>Boulder</td>
<td>Introductory</td>
<td>10</td>
</tr>
<tr>
<td>TC10</td>
<td>1998</td>
<td>Boulder</td>
<td>Introductory</td>
<td>12</td>
</tr>
<tr>
<td>TC11</td>
<td>1998</td>
<td>Leuven</td>
<td>Introductory</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Advanced</td>
<td>13</td>
</tr>
<tr>
<td>TC12</td>
<td>1999</td>
<td>Boulder</td>
<td>Advanced</td>
<td>12</td>
</tr>
<tr>
<td>TC13</td>
<td>2000</td>
<td>Boulder</td>
<td>Introductory</td>
<td>12</td>
</tr>
<tr>
<td>TC14</td>
<td>2001</td>
<td>Boulder</td>
<td>Advanced</td>
<td>18</td>
</tr>
<tr>
<td>TC15</td>
<td>2002</td>
<td>Boulder</td>
<td>Introductory</td>
<td>18</td>
</tr>
<tr>
<td>TC16</td>
<td>2003</td>
<td>Boulder</td>
<td>Advanced</td>
<td>15</td>
</tr>
<tr>
<td>TCE1</td>
<td>2003</td>
<td>Egmond</td>
<td>Introductory</td>
<td>15</td>
</tr>
<tr>
<td>TC17</td>
<td>2004</td>
<td>Boulder</td>
<td>Introductory</td>
<td>18</td>
</tr>
<tr>
<td>TCE2</td>
<td>2004</td>
<td>Egmond</td>
<td>Advanced*</td>
<td>16</td>
</tr>
<tr>
<td>TC18</td>
<td>2005</td>
<td>Boulder</td>
<td>Advanced</td>
<td>18</td>
</tr>
<tr>
<td>TCE3</td>
<td>2005</td>
<td>Egmond</td>
<td>Advanced*</td>
<td>13</td>
</tr>
<tr>
<td>TC19</td>
<td>2006</td>
<td>Boulder</td>
<td>Introductory</td>
<td>15</td>
</tr>
<tr>
<td>TCE4</td>
<td>2006</td>
<td>Egmond</td>
<td>Advanced</td>
<td>12</td>
</tr>
<tr>
<td>TC20</td>
<td>2007</td>
<td>Boulder</td>
<td>Advanced</td>
<td>21</td>
</tr>
<tr>
<td>TC21</td>
<td>2007</td>
<td>Leuven</td>
<td>Anniversary</td>
<td></td>
</tr>
</tbody>
</table>
## Frequency of attendance of faculty and students

<table>
<thead>
<tr>
<th>Frequency</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>16</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Student</td>
<td>585</td>
<td>169</td>
<td>36</td>
<td>14</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>814</td>
</tr>
</tbody>
</table>

- Introductory Workshop # of Students: 920
- Advanced Workshop # of Students: 365
- Total: 1185
Complex Trait Model

- Marker
- Gene\(^1\)
- Gene\(^2\)
- Gene\(^3\)
- Disease Phenotype
- Individual environment
- Common environment
- Polygenic background
- Linkage
- Linkage disequilibrium
- Mode of inheritance

Linkage associations are shown with arrows.
Using genetics to dissect metabolic pathways: Drosophila eye color

Beadle & Ephrussi, 1936
Beadle and Ephrussi, 1936
Finding QTLs

- Linkage
- Association
Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn’s disease

Jean-Pierre Hugot†‡‡, Mathias Chamaillard†‡, Habib Zouali†, Suzanne Lesage†, Jean-Pierre Cézard‡, Jacques Belaiche§, Sven Almer‖, Curt Tyski, Colm A. O’Morain∥, Miquel Gassull‡, Vibeke Binder†∥, Yigael Finkel†‡, Antoine Cortot‡‡, Robert Modigliani§§, Pierre Laurent-Puig†, Corine Gower-Rousseau‡‡, Jeanne Macry¶¶, Jean-Frédéric Colombel‡‡, Mourad Sahbatou† & Gilles Thomas†‡‡

NATURE | VOL 411 | 31 MAY 2001

First (unequivocal) positional cloning of a complex disease QTL!
Linkage analysis
Thomas Hunt Morgan – discoverer of linkage
Linkage = Co-segregation

Marker allele $A_1$ cosegregates with dominant disease
Linkage Markers: microsatellite / SNP/ ...
IDENTITY BY DESCENT

Sib 1

Sib 2

4/16 = 1/4 sibs share BOTH parental alleles  IBD = 2
8/16 = 1/2 sibs share ONE parental allele  IBD = 1
4/16 = 1/4 sibs share NO parental alleles  IBD = 0
For disease traits (affected/unaffected)

Affected sib pairs selected

Markers

Expected

IBD = 1

2BD = 2

1BD = 3

0

1000

750

500

250

127

310
For continuous measures

Unselected sib pairs

<table>
<thead>
<tr>
<th>IBD</th>
<th>Correlation between sibs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Twin 1 mole count

Twin 2 mole count

\[ r_{MZ} = r_{DZ} = 1 \]

\[ r_{MZ} = 1, r_{DZ} = 0.5 \]

\[ r_{MZ} = 1, r_{DZ} = \frac{1}{\pi} \]
Linkage for mole counts in Australian twin families

A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions

Gu Zhu¹, Grant W Montgomery¹, Michael R James¹, Jeff M Trent², Nicholas K Hayward¹, Nicholas G Martin¹ and David L Duffy*.¹

Flat mole count: chromosome 9 linkage in Australian and UK twins

Australia

UK
Linkage for MaxCigs24 in Australia and Finland

AJ HG, in press
log Reading Linkage: Chromosomes 1-22

log Spelling Linkage: Chromosomes 1-22
Effect of multivariate analysis on linkage power

Multivariate and univariate linkage analysis of six reading-related measures on chromosome 18


Use of Multivariate Linkage Analysis for Dissection of a Complex Cognitive Trait

Angela Marlow, Simon Fisher, Clyde Francks, Laurence MacPhie, Stacey Cherny, Alex Richardson, Joel Talcott, John Stein, Anthony Monaco, and Lon Cardon
Robust Estimation of Experimentwise P Values Applied to a Genome Scan of Multiple Asthma Traits Identifies a New Region of Significant Linkage on Chromosome 20q13

Manuel A. R. Ferreira,1 Louise O’Gorman,1 Peter Le Souëf,2 Paul R. Burton,2 Brett G. Toelle,3 Colin F. Robertson,4 Peter M. Visscher,1 Nicholas G. Martin,1 and David L. Duffy1
Ridge count

The size of prints can be measured by counting the number of ridges from the triradii to the core

Ridge count can be summed over all fingers to give a total ridge count

Diagram from Holt, 1968

Highly heritable:

- MZ $r = .94$  CI $0.89 - 0.96$
- DZ $r = .42$  CI $0.34 - 0.50$
- A  $= .82$  CI $0.56 - 0.95$
- D  $= .11$  CI $0.00 - 0.37$
- E  $= .07$  CI $0.05 - 0.10$
TRC vs Multivariate \((-\log_{10} p)\)

\[ \chi^2_1 \quad \chi^2_5 \]
Chromosome 1
Similar ‘drop chi-squares’ for pleiotropic QTLs

Resulting in a very conservative test
Chromosome 7 ...

Evidence of developmental fields?

[Graph showing DNA markers and significant results]
Extreme Discordant Sib Pairs for Mapping Quantitative Trait Loci in Humans

Neil Risch* and Heping Zhang

Analysis of differences between siblings (sib pair analysis) is a standard method of genetic linkage analysis for mapping quantitative trait loci, such as those contributing to hypertension and obesity, in humans. In traditional designs, pairs are selected at random or with one sib having an extreme trait value. The majority of such pairs provide little power to detect linkage; only pairs that are concordant for high values, low values, or extremely discordant pairs (for example, one in the top 10 percent and the other in the bottom 10 percent of the distribution) provide substantial power. Focus on discordant pairs can reduce the amount of genotyping necessary over conventional designs by 10- to 40-fold.
Information Score for Additive Gene Action (p=0.5)
Genotypes available on EDAC plus others

Phenotyped for Neuroticism

Extreme Discordant Concordant Design

Genotyped EDAC plus

# QISPs

<table>
<thead>
<tr>
<th></th>
<th>Neuroticism Phenotyped</th>
<th>QISPs share &gt;300 markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concordant Hi</td>
<td>556</td>
<td>343</td>
</tr>
<tr>
<td>Concordant Lo</td>
<td>717</td>
<td>497</td>
</tr>
<tr>
<td>Discordant</td>
<td>726</td>
<td>463</td>
</tr>
<tr>
<td>The rest</td>
<td>8482</td>
<td>858</td>
</tr>
<tr>
<td>Total</td>
<td>10481</td>
<td>2161</td>
</tr>
</tbody>
</table>
Linkage scan EDAC sample – CIDI interview

Depression traits - OZ
Information for marker density 0.5, 1, 2, 10cM scan

Limits of fine mapping a quantitative trait
Attwood LD & Heard-Costa NL.
Whole-Genome Scan, in a Complex Disease, Using 11,245
Single-Nucleotide Polymorphisms: Comparison with Microsatellites

Sally John,1 Neil Shephard,1 Guoying Liu,2 Eleftheria Zeggini,1 Manqiu Cao,2 Wenwei Chen,2
Nisha Vasavada,3 Tracy Mills,3 Anne Barton,1 Anne Hinks,1 Steve Eyre,1 Keith W. Jones,2
William Ollier,1 Alan Silman,1 Neil Gibson,3 Jane Worthington,1 and Giulia C. Kennedy2

1University of Manchester, Manchester, United Kingdom; 2Affymetrix, Santa Clara, CA; and 3AstraZeneca, Macclesfield, United Kingdom

Despite the theoretical evidence of the utility of single-nucleotide polymorphisms (SNPs) for linkage analysis, no whole-genome scans of a complex disease have yet been published to directly compare SNPs with microsatellites. Here, we describe a whole-genome screen of 157 families with multiple cases of rheumatoid arthritis (RA), performed using 11,245 genomewide SNPs. The results were compared with those from a 10-cM microsatellite scan in the same cohort. The SNP analysis detected HLA$^\text{a}$DRB1, the major RA susceptibility locus ($P = .00004$), with a linkage interval of 31 cM, compared with a 50-cM linkage interval detected by the microsatellite scan. In addition, four loci were detected at a nominal significance level ($P < .05$) in the SNP linkage analysis; these were not observed in the microsatellite scan. We demonstrate that variation in information content was the main factor contributing to observed differences in the two scans, with the SNPs providing significantly higher information content than the microsatellites. Reducing the number of SNPs in the marker set to 3,300 (1-cM spacing) caused several loci to drop below nominal significance levels, suggesting that decreases in information content can have significant effects on linkage results. In contrast, differences in maps employed in the analysis, the low detectable rate of genotyping error, and the presence of moderate linkage disequilibrium between markers did not significantly affect the results. We have demonstrated the utility of a dense SNP map for performing linkage analysis in a late-age-at-onset disease,
Linkage

- Doesn’t depend on “guessing gene”
- Works over broad regions (good for getting in right ball-park) and whole genome (“genome scan”)
- Only detects large effects (>10%)
- Requires large samples (10,000’s?)
- Can’t guarantee close to gene
Association

- Looks for correlation between specific alleles and phenotype (trait value, disease risk)
Association

- More sensitive to small effects
- Need to “guess” gene/alleles ("candidate gene") or be close enough for linkage disequilibrium with nearby loci
- May get spurious association ("stratification") – need to have genetic controls to be convinced
Variation: Single Nucleotide Polymorphisms

Complex disease marker? SNPs are single-base differences in DNA.
Differences (between subjects) in DNA sequence are responsible for (structural) differences in proteins.
Human OCA2 and eye colour

LD blocks in OCA2

Association with eye color
Eye colour explained

A Three–Single-Nucleotide Polymorphism Haplotype in Intron 1 of OCA2 Explains Most Human Eye-Color Variation

David L. Duffy,* Grant W. Montgomery,* Wei Chen, Zhen Zhen Zhao, Lien Le, Michael R. James, Nicholas K. Hayward, Nicholas G. Martin, and Richard A. Sturm

American Journal of Human Genetics  Volume 80   February 2007
Comparison of Affymetrix 10k, 100k, 500k SNP chips

Median intermarker distance: 3.3 kb
Mean intermarker distance: 5.4 kb
Average Heterozygosity: 0.30
Average minor allele frequency: 0.22
SNP Genotyping Platforms

**Throughput (SNPs Per Assay)**

- **TaqMan 7900**: 1?
- **Sequenom MassARRAY**: 25
- **Illumina BeadStation**: 1536

**Cost Per Assay**

**Flexibility in Project Design**
Unprecedented Call Rates of >99%

FROM FUNG AND SINGLETON ET AL. NEUROLOGY THE LANCET

Genome-wide genotyping in Parkinson’s disease and neurologically normal controls: first stage analysis and public release of data


Summary
Background Several genes underlying rare monogenic forms of Parkinson’s disease have been identified over the past decade. Despite evidence for a role for genetics in sporadic Parkinson’s disease, few common genetic variants have been unequivocally linked to this disorder. We sought to identify any common genetic variability exerting a large effect in risk for Parkinson’s disease in a population cohort and to produce publicly available genome-wide genotype data that could be used by others to study genetic variations in Parkinson’s disease.

Methods We performed genome-wide genotypes in 537 individuals, 272 of whom were patients and 265 were age-matched controls. We used a composite Illumina BeadChip panel that included approximately 550,000 SNPs.

Findings We have produced around 220 million genotypes in 537 participants. This raw genotype data has been publicly posted and as such is the first publicly accessible high-density SNP data outside of the International HapMap Project. We also provide here the results of genotype and allele association tests.

Interpretation We generated publicly available genotype data for Parkinson’s disease patients and controls so that these data can be mined and augmented by other researchers to identify common genetic variability that results in minor and moderate risk for disease.

A total of 219,577,497 unique genotype calls were made and the average call rate across all samples was 99.6%.
GWAS for Inflammatory Bowel Disease

Frequency distribution of SNP association tests in a genome scan

- Noise: Irrelevant SNPs
- Gene variants related to disease

Test statistic

Known CD gene
CARD15

Novel CD gene
IL23R

Three exceptional SNPs (+)

Re-examine scan data for other loci in IL23R

Additional disease-associated variants discovered (■)

Delivering New Disease Genes
Lon R. Cardon
GWAS for Inflammatory Bowel Disease

A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene
A genome-wide association study identifies novel risk loci for type 2 diabetes

Stage 1: Illumina 100k+300k
Stage 2: Sequenom Iplex
A genome-wide association study identifies novel risk loci for type 2 diabetes

Philipppe Bouchard†, Daniel Vincent†, Alexandre Belisle†, Sammy Hadjadj†, Beverley Balkau†, Barbara Heude†,
Guillaume Charpentier†, Thomas J. Hudson**, Alexandre Montpetit†, Alexey V. Pechersky**, Marc Forget†,†,†,
Barry I. Posner†,†, David J. Balding†, David Meyre†, Constantine Polychronakos† & Philippe Froguel†,††

*Equal contribution
**Lead investigator
†Corresponding author
Cutting costs of GWAS by DNA pooling

Case DNAs

| DNA1 | DNA2 | .......... | DNA300 |

Control DNAs

| DNA1 | DNA2 | .......... | DNA300 |

case pool (N=384)

control pool (N=384)

Affymetrix Genechip Hind III arrays
Common KIBRA Alleles Are Associated with Human Memory Performance

Fig. 3. Significant KIBRA allele–dependent differences in hippocampal activation as measured with fMRI. Activations are significantly increased in the hippocampus in noncarriers (n = 15) of the T allele.
Pooling error for 15,000 SNPs using Illumina Hap300 and Affy 50k arrays

Illumina arrays extract 80% information as IG vs ~30% with Affy: need ~10x Affy arrays

Stuart Macgregor, QIMR
Illumina Hap300 versus Affy 50k array-specific error plots

Illumina 317k data

Aftymetrix 50K HindIII data

Stuart Macgregor, QIMR
Role of miRNA (binding sites) in disease?

Sequence Variants in *SLITRK1*
Are Associated with Tourette’s Syndrome
A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep and in quantitative traits.

Texel sheep are renowned for their exceptional meatiness. To identify the genes underlying this economically important feature, we performed a whole-genome scan in a Romanov × Texel F2 population. We mapped a quantitative trait locus with a major effect on muscle mass to chromosome 2 and subsequently fine-mapped it to a chromosome interval encompassing the myostatin (GDF8) gene. We herein demonstrate that the GDF8 allele of Texel sheep is characterized by a G to A transition in the 3’ UTR that creates a target site for mir1 and mir206, microRNAs (miRNAs) that are highly expressed in skeletal muscle. This causes translational inhibition of the myostatin gene and hence contributes to the muscular hypertrophy of Texel sheep. Analysis of SNP databases for humans and mice demonstrates that mutations creating or destroying putative miRNA target sites are abundant and might be important effectors of phenotypic variation.
Even for “simple” diseases the number of alleles is large

- Ischaemic heart disease (LDR) >190
- Breast cancer (BRAC1) >300
- Colorectal cancer (MLN1) >140
Multiple Rare Alleles Contribute to Low Plasma Levels of HDL Cholesterol

Jonathan C. Cohen,1,2,3,4 Robert S. Kiss,5,6
Alexander Pertseremidis,1 Yves L. Marcel,5 Ruth McPherson,5
Helen H. Hobbs1,3,4

Heritable variation in complex traits is generally considered to be conferred by common DNA sequence polymorphisms. We tested whether rare DNA sequence variants collectively contribute to variation in plasma levels of high-density lipoprotein cholesterol (HDL-C). We sequenced three candidate genes (ABCA1, APOA1, and LCAT) that cause Mendelian forms of low HDL-C levels in individuals from a population-based study. Nonsynonymous sequence variants were significantly more common (16% versus 2%) in individuals with low HDL-C (<fifth percentile) than in those with high HDL-C (>95th percentile). Similar findings were obtained in an independent population, and biochemical studies indicated that most sequence variants in the low HDL-C group were functionally important. Thus, rare alleles with major phenotypic effects contribute significantly to low plasma HDL-C levels in the general population.

Complex disease: common or rare alleles?

Increasing evidence for Common Disease – Rare Variant hypothesis (CDRV)

Table 1. Sequence variations in the coding regions of ABCA1, APOA1, and LCAT. Values represent the numbers of sequence variants identified in 256 individuals from the Dallas Heart Study (DHS) (128 with low HDL-C and 128 with high HDL-C) and 263 Canadians (155 with low HDL-C and 108 with high HDL-C) (17). NS, nonsynonymous (nucleotide substitutions resulting in an amino acid change); S, synonymous (coding sequence substitutions that do not result in an amino acid change). GenBank accession numbers for DHS ABCA1, APOA1, and LCAT sequences are NM_005502, NM_000039, and NM_000229, respectively.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sequence variants unique to one group</th>
<th>Sequence variants common to both groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low HDL-C</td>
<td>High HDL-C</td>
</tr>
<tr>
<td>ABCA1</td>
<td>14 6</td>
<td>2 5</td>
</tr>
<tr>
<td>APOA1</td>
<td>1 0</td>
<td>0 1</td>
</tr>
<tr>
<td>LCAT</td>
<td>0 1</td>
<td>1 0</td>
</tr>
</tbody>
</table>

DHS

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sequence variants unique to one group</th>
<th>Sequence variants common to both groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low HDL-C</td>
<td>High HDL-C</td>
</tr>
<tr>
<td>ABCA1</td>
<td>14 2</td>
<td>2 3</td>
</tr>
<tr>
<td>APOA1</td>
<td>0 1</td>
<td>0 0</td>
</tr>
<tr>
<td>LCAT</td>
<td>6 1</td>
<td>0 0</td>
</tr>
</tbody>
</table>

Canadians

[Science 2004]
Product Portfolio and Application Areas

- **Human-1**: Gene centric association studies
- **Hap300**: WGA with 80% genomic coverage in Caucasians
- **Hap550**: WGA with ~90% genomic coverage in Caucasians and Asians
- **Hap650Y**: WGA with ~90% genomic coverage in Caucasians and Asians AND 67% genomic coverage in Yoruban

**Hap450S + Hap240S**

**1M**: The most comprehensive chip that allows whole genome DNA analysis with industry leading SNP coverage in genes, CNV regions and indels
## 1M Content

<table>
<thead>
<tr>
<th>CONTENT</th>
<th>NUMBER</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanHap550</td>
<td>555,000</td>
<td>High genomic coverage</td>
</tr>
<tr>
<td>SNPs in Genes</td>
<td>400,000</td>
<td>High density of SNPs in coding regions of the genome</td>
</tr>
<tr>
<td>SNPs and Probes in both reported and novel Copy Number Variant (CNV) Regions</td>
<td>110,000</td>
<td>High density of SNPs and probes in CNV regions, including “nonSNPable” regions</td>
</tr>
<tr>
<td>Additional Caucasian and Asian Tag SNPs</td>
<td>84,000</td>
<td>Higher tag SNP coverage of the genome</td>
</tr>
<tr>
<td>Additional African Tag SNPs</td>
<td>100,000</td>
<td>Higher tag SNP coverage of the genome</td>
</tr>
<tr>
<td>Even Spacing SNPs</td>
<td>90,000</td>
<td>Ensure complete coverage across the genome, enable new CNV discoveries</td>
</tr>
<tr>
<td>ADME/MHC SNPs</td>
<td>17,000</td>
<td>Denser coverage in high value regions/genes</td>
</tr>
<tr>
<td>TOTAL</td>
<td>&gt;1M</td>
<td>Unsurpassed power and gene coverage for WGA and CNV studies</td>
</tr>
</tbody>
</table>
Human 1M HapMap Coverage by Population

GENOME COVERAGE ESTIMATED FROM 990,000 HAPMAP SNPs IN HUMAN 1M

~95%
~94%
~74%
Copy Number Variation (CNV) in MZ twin pair
The $1,000 Human Genome - Implications for Life Science, Healthcare, and IT

Published by:
IDC
solexa sequencing applications

Illumina's Solexa Sequencing technology offers a powerful new approach to some of today's most important applications for genetic analysis and functional genomics, including:

sequencing and resequencing
Whether you need to sequence an entire genome or a large candidate region, the Illumina Genome Analyzer System is today's most productive and economical sequencing tool. Solexa sequencing technology and reversible terminator chemistry deliver unprecedented volumes of high quality data, rapidly and economically.

expression profiling
Sequencing millions of short cDNA tags per sample, the Genome Analyzer allows you to generate digital expression profiles at costs comparable to current analog methods. Because our protocol does not require any transcript-specific probes, you can apply the technology to discover and quantitate transcripts in any organisms, irrespective of the annotation available on the organism.

small rna identification and quantification
Solexa sequencing technology also offers a unique and powerful solution for the comprehensive discovery and characterization of small RNAs in a wide range of species. The massively parallel sequencing protocol allows researchers to discover and analyze genome-wide profiles of small RNA in any species. With the potential to generate several million sequence tags economically, the Illumina Genome Analyzer offers investigators the opportunity to uncover global profiles of small RNA at an unprecedented scale.
EPIGENETIC DISCORDANCE IN IDENTICAL TWINS

The missing “environment”? 
urethra, a dilated pelvis of the right kidney, bilateral uterus unicornis with normal ovaries, hemivertebrae of thoracic vertebrae 6 and 10, and abnormal curvature of the sacrum. A persistent ductus arteriosus and secundum atrial septum defect was suspected, but results of cardiac investigations at 10 months were normal.

At physical examination for genetic evaluation at 4 months we saw a baby girl with epicanthal folds, but no other minor anomalies. She had a capillary nevus on her left buttock. In the anal region only a dimple was seen. The patient was operated on one day after birth, when a colostomy was made and a fistula connected to the colon descended. At age 10 months she 8amina and left
Discordant caudal duplication in MZ twins

Twin 1 - unaffected

Twin 2 - affected

Controls [e.g.]

Emma Whitelaw, Suyinn Chong
Department of Biochemistry
University of Sydney
Other studies on MZ discordance

**Epilepsy**  (with S. Berkovic, L. Vadlamudi)

**Schizophrenia**  (with B. Mowry, N. Hayward)

**Depression**  (with A. Petronis, D. Boomsma, P. McGuffin)

**Asthma**  (with M. Ferreira, E. Whitelaw)
We also run two journals (1)

- Editor: John Hewitt
- Editorial assistant: Christina Hewitt
- Publisher: Kluwer/Plenum
- Fully online
- http://www.bga.org
We also run two journals (2)

- Editor: Nick Martin
- Editorial assistant + subscriptions: Marisa Grimmer
- Publisher: Australian Academic Press
- Fully online
Twin 1 phenotype

E

D

A

Q

Twin 1 phenotype

E

D

A

Q

Twin 2 phenotype

r_{MZ} = 1, r_{DZ} = 0.25

r_{MZ} = 1, r_{DZ} = 0.5

r_{MZ} = 1, r_{DZ} = \frac{1}{\pi}
But why do we use the average sib values of
\[ r_a = 0.5 \]
\[ r_d = 0.25 \]
when we can estimate the (almost) exact values for each sib pair from marker data?

Are there any advantages in doing so?
Mean IBD sharing across the genome for the $j$th sib pair was based on IBD estimated from Merlin every centimorgan and averaged at all 3491 points

**additive**

$$\overline{\hat{\pi}}_{a(j)} = \frac{\sum_{i=1}^{3491} \hat{\pi}_{a(ij)}}{3491}$$

**dominance**

$$\overline{\hat{\pi}}_{d(j)} = \frac{\sum_{i=1}^{3491} p_{2(ij)}}{3491}$$
Application

• Phenotype = height

Number of sibpairs with phenotypes and genotypes

*Adolescent cohort* 931
*Adult cohort* 2444
*Combined* 3375
Mean and SD of genome-wide additive relationships

Mean = 0.4984
Std. Dev. = 0.03598
N = 4,401
Mean and SD of genome-wide dominance relationships

**Histogram:**
- **Frequency**
- **X-axis:** ibd2_mean
- **Y-axis:** Frequency

**Data Points:**
- **Mean:** 0.2479
- **Std. Dev.:** 0.04001
- **N:** 4,401
Additive and dominance relationships
correlation = 0.91 (n= 4401)
Models

F = Family effect
A = Genome-wide additive genetic
E = Residual

Full model  \[ F + \hat{\tau}_{a(j)}A + E \]
Reduced model  \[ F + E \]
# Sampling variances are large

<table>
<thead>
<tr>
<th>Cohort</th>
<th>F+A (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adolescent</td>
<td>0.80 (0.36 – 0.90)</td>
</tr>
<tr>
<td>Adult</td>
<td>0.80 (0.61 – 0.86)</td>
</tr>
<tr>
<td>Combined</td>
<td>0.80 (0.62 – 0.85)</td>
</tr>
</tbody>
</table>

*Estimates of MZ correlation from fullsibs!*  

PLOS Genetics, *in press*

And now for IQ! Anyone got sibpairs with IQ + genome scan?
Comparative Genomics

= differences in DNA sequence

Human-Human 1:1000 = 0.1%

Human-Chimp 1:100 = 1%

Human-Mouse 1:8 = 15%
Expression profiling in primates reveals a rapid evolution of human transcription factors

Yoav Gilad1, Alicia Oshlack2, Gordon K. Smyth3, Terence P. Speed1,4 & Kevin P. White1

features that point to the action of directional selection. Among the gene set with a human-specific increase in expression, there is an excess of transcription factors; the same is not true for genes with increased expression in chimpanzee.
Which genes have evolved fastest?

Figure 3. Plots of Chromosome 2 SNPs with Extreme iHS Values Indicate Discrete Clusters of Signals.
### Table 1. Summary of Some of the Strongest iHS Signals Genome-Wide

| Cytological Position | Genes (Number) | Size (kb) | Pop  | Number of SNPs with $|iHS| > 2.0$ |
|----------------------|----------------|-----------|------|----------------------------------|
| 1p34.3               | NCDN, TEKT2 (17) | 1,200     | CEU  | 74/103                           |
| 1p31.1               | SLC44A5 (4)     | 900       | ASN  | 97/150                           |
| 2p23.3               | NCOA1, ADCY3 (4) | 400       | YRI  | 51/76                            |
| 2q12.3-q13           | SULT1C cluster (13) | 1,100     | ASN  | 108/171                          |
| 2q21.3-q22.1         | LCT (15)        | 2,800     | CEU  | 351/594                          |
| 2q32.3               | None (0)        | 400       | YRI  | 100/131                          |
| 4p15.1               | None (0)        | 500       | CEU  | 91/125                           |
| 4q21–23              | ADH cluster (8) | 100       | ASN  | 21/28                            |
| 8q11.21–23           | SNTG1 (8)       | 3,100     | ASN  | 129/1297                         |
| 9p22.3               | C9orf93(1)      | 400       | ASN  | 142/204                          |
| 12q21.2              | SYT1 (3)        | 700       | YRI  | 108/143                          |
| 20cen                | ITGB4BP, CEP2, SPAG4 (24) | 800 | ASN  | 101/135                          |
|                      |                 |           | CEU  | 50/153                           |
|                      |                 |           | YRI  | 22/154                           |

---

**A Map of Recent Positive Selection in the Human Genome**

Benjamin F. Voight*, Sridhar Kudaravalli*, Xiaoquan Wen, Jonathan K. Pritchard*
Migraine - Genome Scan Results

MERLIN-regress (LCA 2-class affection)

Chr Boundaries
LOD-regress

LOD vs. Chr Boundaries for MERLIN-deviates vs. MERLIN-regress (LCA 2-class affection). Peaks at Chr 1, 5, 8, 10, and 13.
MERLIN-deviates IHS Symptom Analyses

PHONOPHOBIA  MA, IHS…

ATP1A2 (FHM2)

1q31
MOA
FHM?