Mutation detection, SNP analysis and genetic linkage

Denis Shields

Denis.shields@ucd.ie

Overview

- Forward versus reverse genetics
  = association versus linkage
- Linkage
- Association
- Mutation detection
- Disequilibrium
- Genotyping studies
- Genotyping methods
- Analysis and designing studies

dbSNP

www.hapmap.org

Summary

- **DMMC Course: TECHNIQUES & STRATEGIES IN MOLECULAR MEDICINE**
- 1130-1215 Monday 10 December 2007, Panoz institute, LTEE2, TCD
- **Detecting common and rare genetic variants associated with disease.** Prof Denis Shields (UCD Conway Institute of Biomolecular & Biomedical Research)
- Different strategies are required to identify rare and common genetic variants underlying both rare and common diseases. For common genetic variants, there is now a very rich dataset of identified common single nucleotide polymorphisms (SNPs). These can be investigated in disease groups (compared to controls) in candidate genes, or by whole genome association analysis. Analysis of these genes requires careful attention to the patterns of association of SNPs that are chromosomally adjacent (in linkage disequilibrium). Linkage analysis (tracking in families the disease co-inheritance with widely spaced gene markers) is the traditional approach of choice for rare mutations that have strong phenotypic effects. High throughput sequencing of candidate regions (and in future whole genomes) are accelerating the rate of data accumulation.
- References: Nature 429:446

Forward or reverse genetics?

- FORWARD
  - Candidate genes
  - Look to see what effect variants in the gene have
  - “Association study”
- REVERSE
  - Start with a phenotype that looks genetic
  - Look to see what regions of the human genome travel in families with the phenotype
  - “Linkage study”

- WHOLE GENOME ASSOCIATION
  - Every region is a candidate region
  - Look to see what effect variants in the gene have
  - “Whole genome association study”

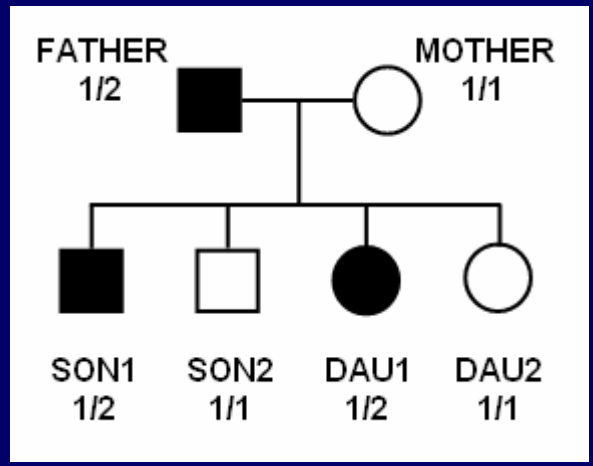
Linkage

- The probability of recombination in families always increases between markers as move further along the chromosome.
- Need a good number of markers spread across the genome (a few hundred) to have a chance of finding a marker-disease association.
- Need a good number of affected in families in order to map the gene to a narrow section of chromosome
- Can use RFLPs (dinucleotide/trinucleotide microsatellites) or recently SNPs (less error in calling genotype).
  - RFLP=Restriction Fragment Length Polymorphism
  - SNP = Single nucleotide polymorphism

Linkage and disease genes

- Recombination fraction (theta) between M & D
  - probability of being passed down together in meiosis
  - Minimum 0, maximum 0.5.
  - Morgan = probability of a recombination happening
    - Allows for chance of > 1 recombination
  - Genetic maps are measured in centimorgans (cM)
  - Each chromosome has between 1 and a few cross-overs per generation (large chromosomes, more cross-overs)

  - Basis of linkage studies:
    - D: unknown disease gene,
    - M: is a marker gene,
      - Then marker gene travels in families with the disease phenotype.



Linkage versus association

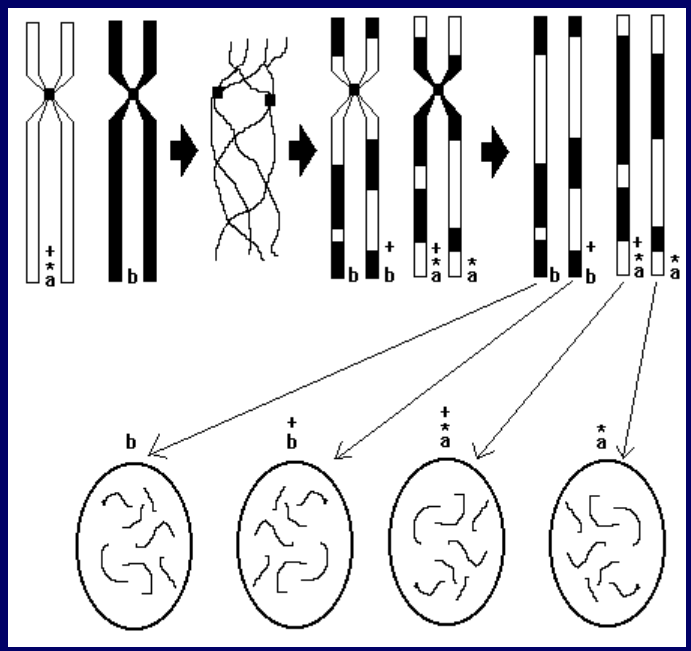
- Linkage
  - Family studies trace linkage of widely spaced markers to disease genes
  - works well for rarer traits running in families
    - familial lipidemias
    - diabetes
  - Not good for common traits with no single genetic cause
    - hypertension
- Association
  - Compares gene frequencies between groups of people
  - Assessing the role of a polymorphism
    - or a chromosomally adjacent variant which is in “disequilibrium” at population level
  - Better when underlying factors confer a weak risk.
  - Samples from other non-genetic studies often useful

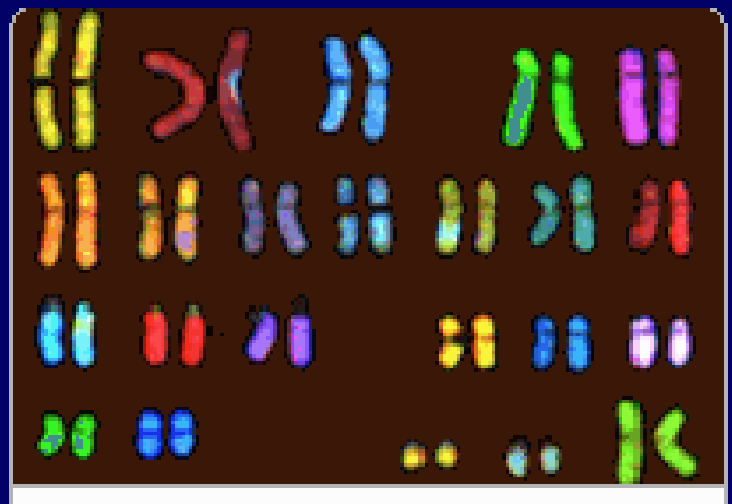
Linkage versus association

- Linkage: increasingly being applied to more detailed intermediate phenotypes in complex disease, rather than just to the ultimate complex outcome
- Risch and Merikangas (1996): power calculations indicate that large association studies may be better than linkage for complex disease outcomes.

Linkage: tools

- Simulation programs to decide study power
  - How many families of what type needed
- Likelihood analysis
    - (programs: Genehunter, Linkage)
  - To correct for multiple regions of human genome linkage scanned, to have the equivalent of a 0.05 chance of inferring linkage when there is none, use a cut off of 1000:1 odds for any single region (LOD=3).





Mutation detection

- Finding new variants: technology
  - Very old days: from studies of protein function
  - Old days: sizing DNA found length variants
    - RFLPs
  - 2 years ago days:
    - Sequencing
      - » Cloning and sequencing
      - » Sequencing human DNA, detecting heterozygote peaks
    - Denaturing HPLC to distinguish heteroduplexes in heterozygote samples.
    - Other methods: Denaturing Gel Electrophoresis
  - Now: massively cheap large scale sequencers
    - Solexa
    - 454
    - ABI solid

Mutations/variants: where to look?

- Finding new variants: strategies
  - First database to inspect: www.hapmap.org
    - millions of SNPs typed
    - Clean database, most SNPs are real
    - Allele frequencies in Caucasian, African, Asian
    - Pattern of relationship among SNPs is clear
      - Which tend to be associated with other SNPs
  - dbSNP and similar databases
      - Are riddled with error but have loads of real data too.
  - Literature: including:
      - Mutation studies detect common SNPs when looking for severe variants.
  - Coding (mRNA)
    - EST databases
  - Upstream
    - Decide strategy on the basis of any experimental data about promoters, conservation between species of DNA regions.
  - Intron:
    - Intron-exon boundaries
    - Regions conserved between species.

Mutation detection

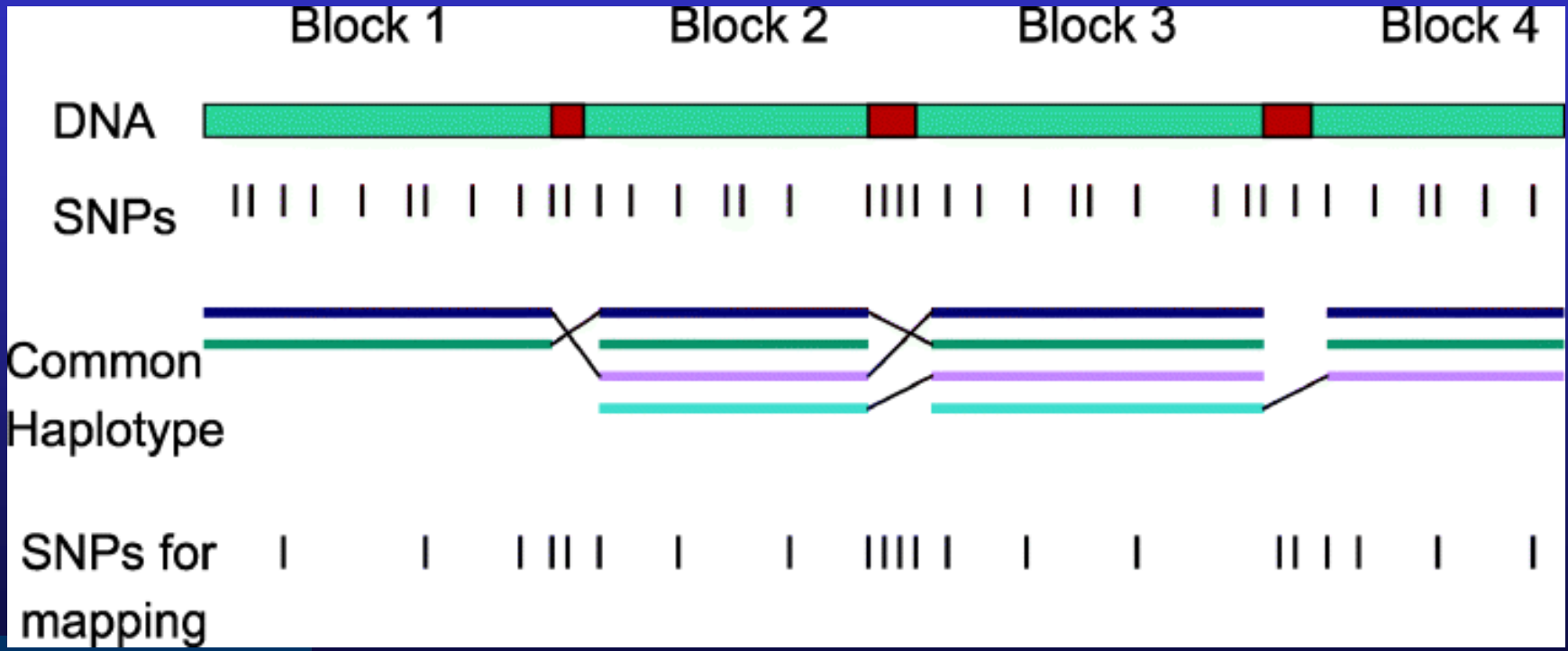
- Should you look for:
  - common variants (just study a small number of people)
  - Rare variants (study a large number, perhaps use pooling).
- This depends on what effects you are looking for
- Are rare alleles or common alleles of more interest in your study?
  - Rare disorder, probably Mendelian
    - Look for variants in the people with the rare characteristic
      - Often in candidate genes from regions of the genome
        - » identified by linkage analysis of the family from which the sample came
  - Common disease
    - Look for variants in some of your own patients
    - Rely on existing databases/literature to get a good coverage

Disequilibrium

- Because genes are linked along chromosomes:
  - Very close genetic variants often travel together at a population level.
  - There have not been enough recombination events since birth of variants
  - Close: e.g. within 100 kb.

Disequilibrium

- Disequilibrium
  - Is not constant along the chromosome
    - Recombinational hotspots.
    - Potential selective sweeps.
    - Is not always greatest for variants nearest to each-other
      - Since the pattern of disequilibrium is very old
        - » Mutations influence what happens.
        - » A recent mutation followed by drift can bring two distant variants into tight disequilibrium
        - » Gene conversion mutations can swap bits among haplotypes in an irregular way.


Block 1
Block 2
Block 3
Block 4
DNA
SNPs
Common Haplotype
SNPs for mapping

HapMap Data (Rel#12 October 04): chr2:190041641..190080014 - Microsoft Internet Explorer provided by RCSI

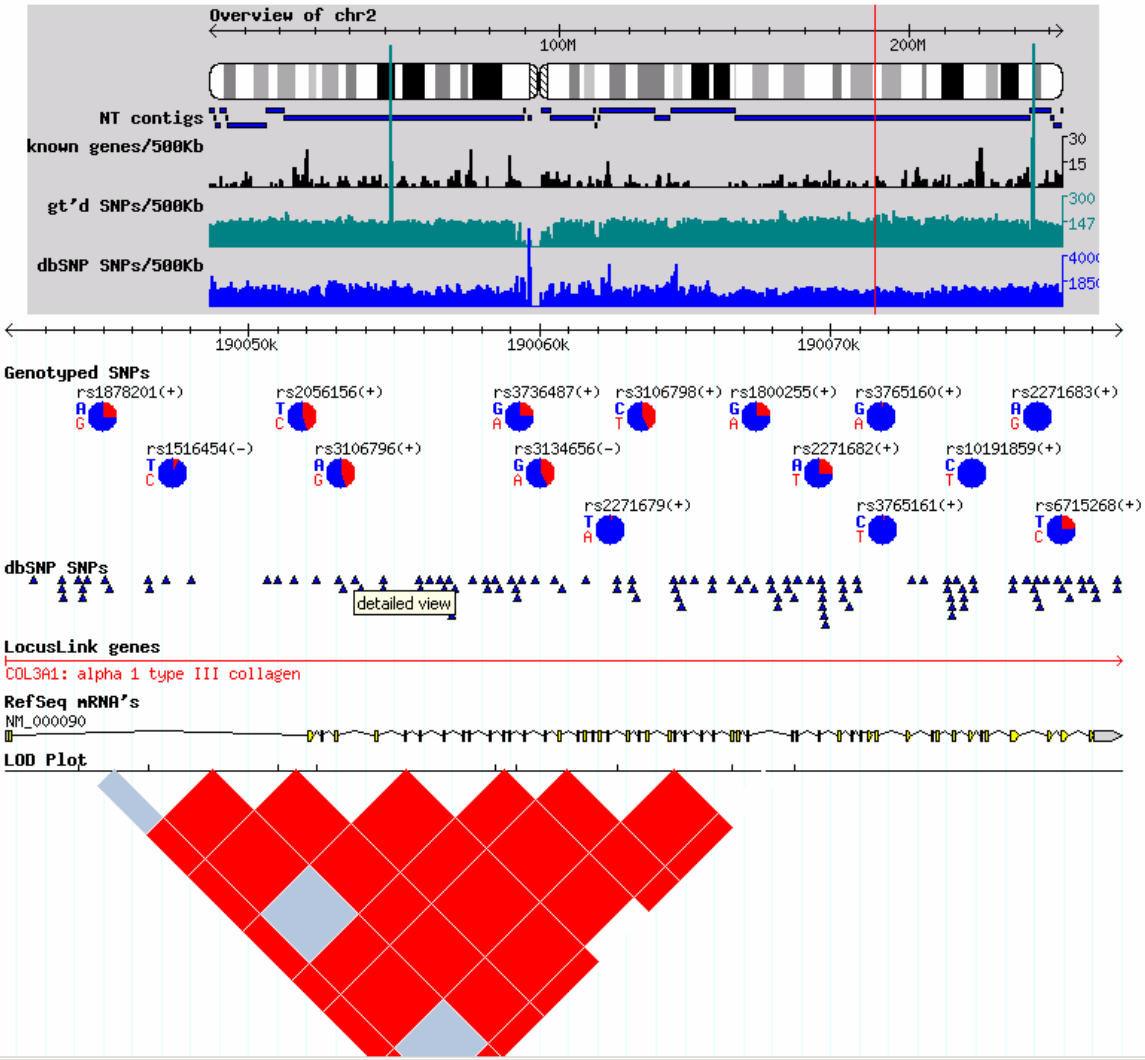
File Edit View Favorites Tools Help

Back Search Favorites History

Address http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap

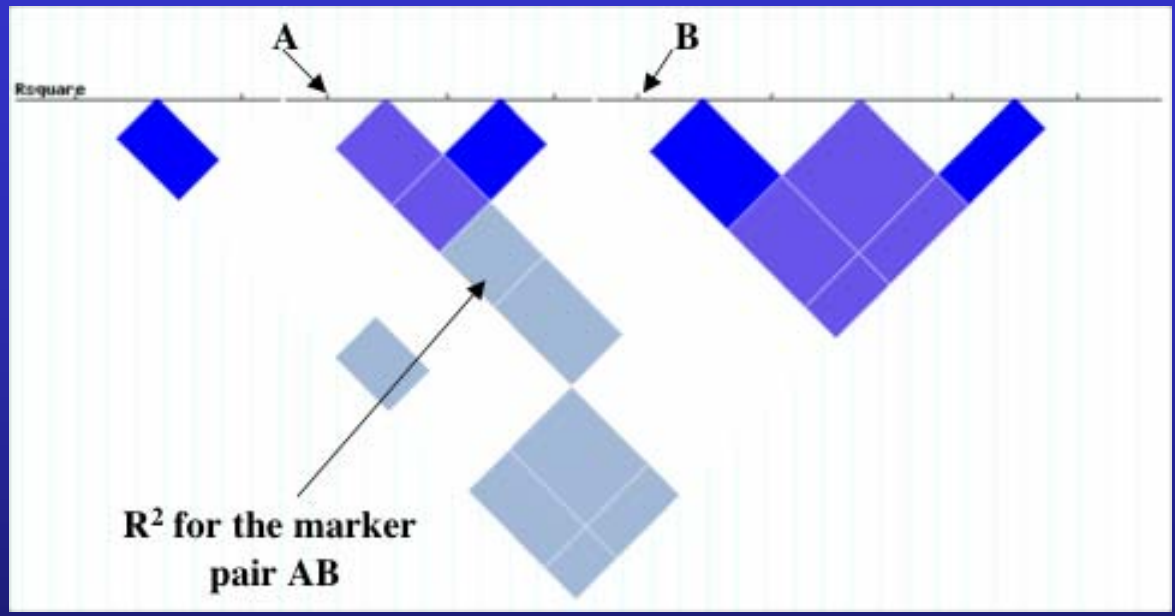
Google Search Web PageRank AutoFill Options

For performing in depth LD and Haplotype analysis of genotype data install Haploview in your local machine
Haploview (ver2.05) is now available for download.

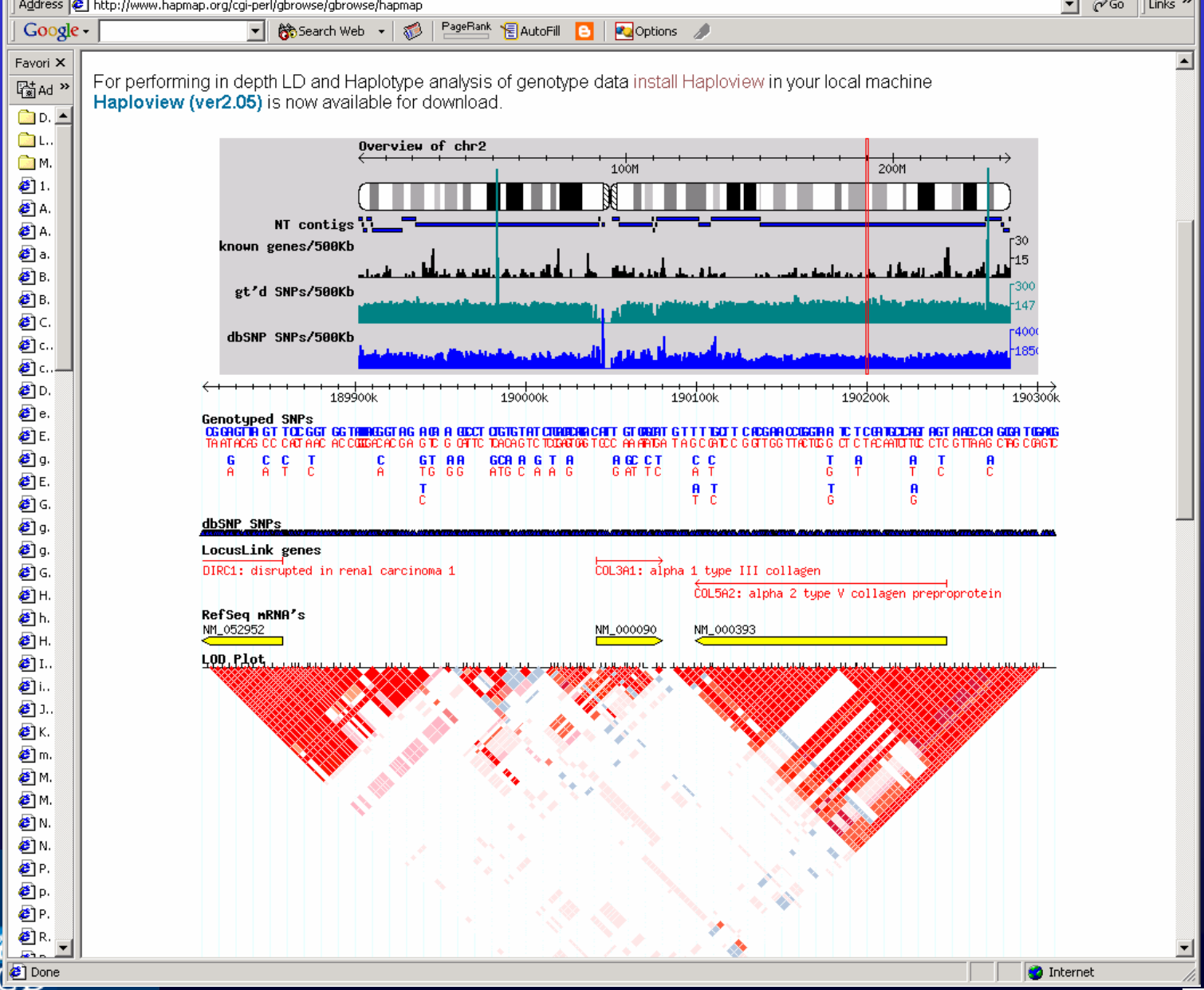


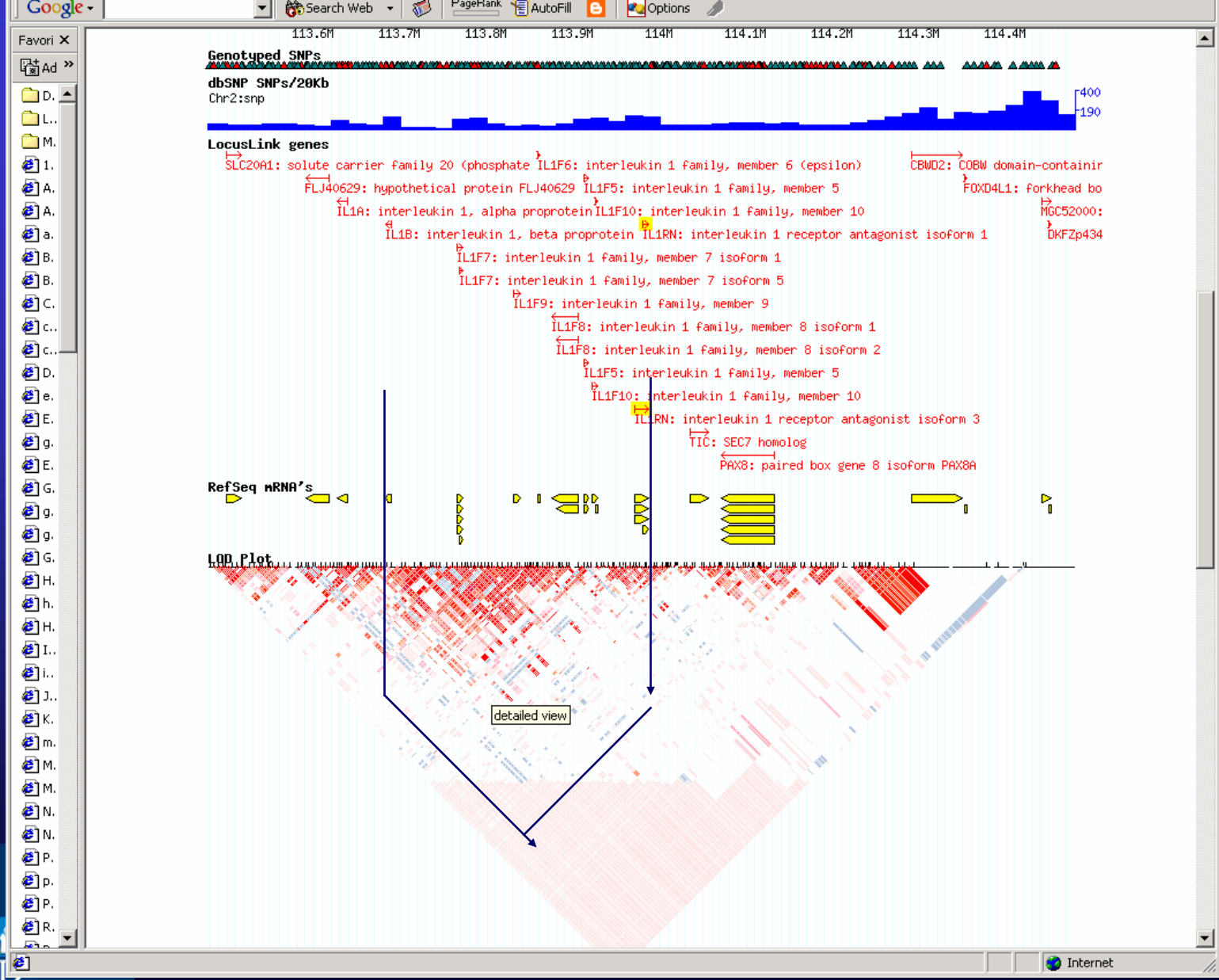
Internet

www.hapmap.org



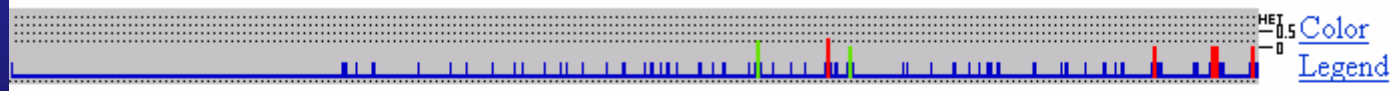
- R-squared: a measure of how much two SNPs travel together
  - Value of 1 completely associated
  - Value of 0, completely unassociated


Address http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap
Google
Search Web
PageRank
AutoFill
Options
Favori
For performing in depth LD and Haplotype analysis of genotype data install Haploview in your local machine
Haploview (ver2.05) is now available for download.
Overview of chr2
100M
200M
NT contigs
known genes/500Kb
gt'd SNPs/500Kb
dbSNP SNPs/500Kb
189900k
190000k
190100k
190200k
190300k
Genotyped SNPs
dbSNP SNPs
LocusLink genes
DIRC1: disrupted in renal carcinoma 1
COL3A1: alpha 1 type III collagen
COL5A2: alpha 2 type V collagen preproprotein
RefSeq mRNA's
NM_052952
NM_000090
NM_000393
LOD Plot
Done
Internet


Google
Search Web
PageRank
AutoFill
Options
Favori
Ad
113.6M
113.7M
113.8M
113.9M
114M
114.1M
114.2M
114.3M
114.4M
Genotyped SNPs
dbSNP SNPs/20Kb
Chr2:snp
400
190
LocusLink genes
SLC20A1: solute carrier family 20 (phosphate
IL1F6: interleukin 1 family, member 6 (epsilon)
CBWD2: COBW domain-containir
FLJ40629: hypothetical protein FLJ40629
IL1F5: interleukin 1 family, member 5
FOXD4L1: forkhead bo
IL1A: interleukin 1, alpha proprotein
IL1F10: interleukin 1 family, member 10
MGC52000:
IL1B: interleukin 1, beta proprotein
IL1RN: interleukin 1 receptor antagonist isoform 1
DKFZp434
IL1F7: interleukin 1 family, member 7 isoform 1
IL1F7: interleukin 1 family, member 7 isoform 5
IL1F9: interleukin 1 family, member 9
IL1F8: interleukin 1 family, member 8 isoform 1
IL1F8: interleukin 1 family, member 8 isoform 2
IL1F5: interleukin 1 family, member 5
IL1F10: interleukin 1 family, member 10
IL1RN: interleukin 1 receptor antagonist isoform 3
TIC: SEC7 homolog
PAX8: paired box gene 8 isoform PAX8A
RefSeq mRNA's
LOD Plot
detailed view
Internet

- dbSNP keeps changing access front end
- Try this:
- www.ncbi.nlm.nih.gov/entrez
  - Select “Gene” as category to search on
  - Enter name of gene you already know
    - (try HUGO gene nomenclature website to get gene names)
  - Click on the human gene that you want
  - From the right panel click “SNP gene view”

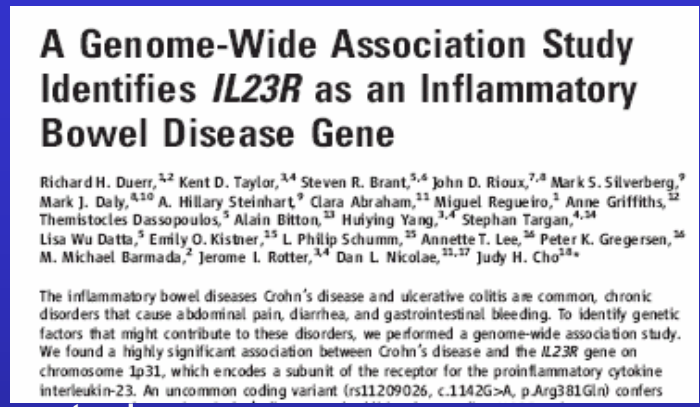
gene model (contig mRNA transcript):	Contig	mrna	protein	mrna orientation	transcript	snp count
	NT_005403	NM_000090	NP_000081	forward	plus strand	7, coding



Contig position	dbSNP rs# cluster id	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino aci position
40071513	rs7579903	0.229				synonymous	A	Gln [Q]	3	617
		0.229				contig reference	G	Gln [Q]	3	617
40073496	rs1800255	0.302	H			nonsynonymous	A	Thr [T]	1	698
		0.302	H			contig reference	G	Ala [A]	1	698
40073998	rs1801184	N.D.				synonymous	C	Gly [G]	3	748
		N.D.				contig reference	T	Gly [G]	3	748
40083153	rs2271683	0.045	H			nonsynonymous	G	Val [V]	1	1205
		0.045	H			contig reference	A	Ile [I]	1	1205
40084837	rs1516446	N.D.	H			nonsynonymous	G	Gln [Q]	3	1353
		N.D.	H			contig reference	T	His [H]	3	1353
40084837	rs17856794	N.D.				nonsynonymous	G	Gln [Q]	3	1353

Cutting down the number of SNPs to analyse:

- Choose a set of SNPs that:
  - maximally represents variation, for minimum cost
  - There are a number of pre-computed sets of SNPs for all genes that will work fine in Caucasian populations



A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Duerr,[1,2] Kent D. Taylor,[3,4] Steven R. Brant,[5,6] John D. Rioux,[7,8] Mark S. Silverberg,[9] Mark J. Daly,[8,10] A. Hillary Steinhart,[9] Clara Abraham,[11] Miguel Regueiro,[1] Anne Griffiths,[12] Themistocles Dassopoulos,[5] Alain Bitton,[13] Huiying Yang,[3,4] Stephan Targan,[4,14] Lisa Wu Datta,[5] Emily O. Kistner,[15] L. Philip Schumm,[15] Annette T. Lee,[16] Peter K. Gregersen,[16] M. Michael Barmada,[2] Jerome I. Rotter,[3,4] Dan L. Nicolae,[11,17] Judy H. Cho[1,8]*

The inflammatory bowel diseases Crohn's disease and ulcerative colitis are common, chronic disorders that cause abdominal pain, diarrhea, and gastrointestinal bleeding. To identify genetic factors that might contribute to these disorders, we performed a genome-wide association study. We found a highly significant association between Crohn's disease and the *IL23R* gene on chromosome 1p31, which encodes a subunit of the receptor for the proinflammatory cytokine interleukin-23. An uncommon coding variant (rs11209026, c.1142G>A, p.Arg381Gln) confers

- Science 314:1461 (2006)
- >500 cases, ileal Crohn's disease >500 controls
- 300,000 SNPs (Illumina technology)
- 2 markers with high p-values
  - rs2066843 (P = 3 × 10−9, corrected P=9×10−4)
  - rs2076756 (P=5 ×10−10,corrected P= 1 × 10−4)
  - are in the known CD gene, CARD15
- Novel association in Interleukin 23
  - rs11209026 (P = 5 × 10−9, corrected P = 1 × 10−3), nonsynonymous SNP (1142G>A, Arg381Gln)

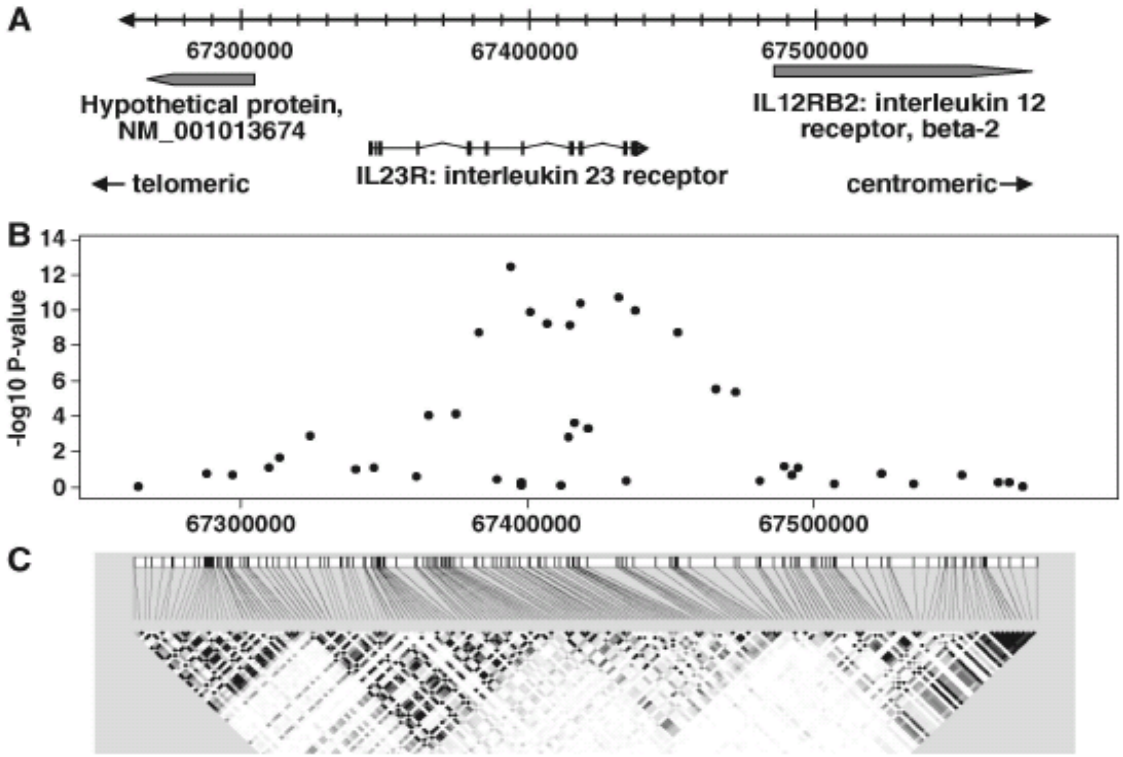


Fig. 1. Association signals in the *IL23R* gene region on chromosome 1p31. (**A**) Genomic locations of genes on chromosome 1p31 between 67,260,000 and 67,580,000 base pairs (Build 35). (**B**) The negative $\log_{10}$ association *P*-values (Cochran-Mantel-Haenszel chi-square test) from the combined Jewish and non-Jewish case-control cohorts are plotted for genotyped markers in the region. (**C**) Pairwise r^2 plot for International HapMap CEU data. The intensity of the shading is proportional to r^2. The *IL23R* gene is contained within two blocks of linkage disequilibrium, and the association signals are strongest in the centromeric block, which contains exons 5 to 11 and extends into the intergenic region between *IL23R* and *IL12RB2*. Note that markers in the block encompassing the *IL12RB2* gene do not demonstrate significant association.

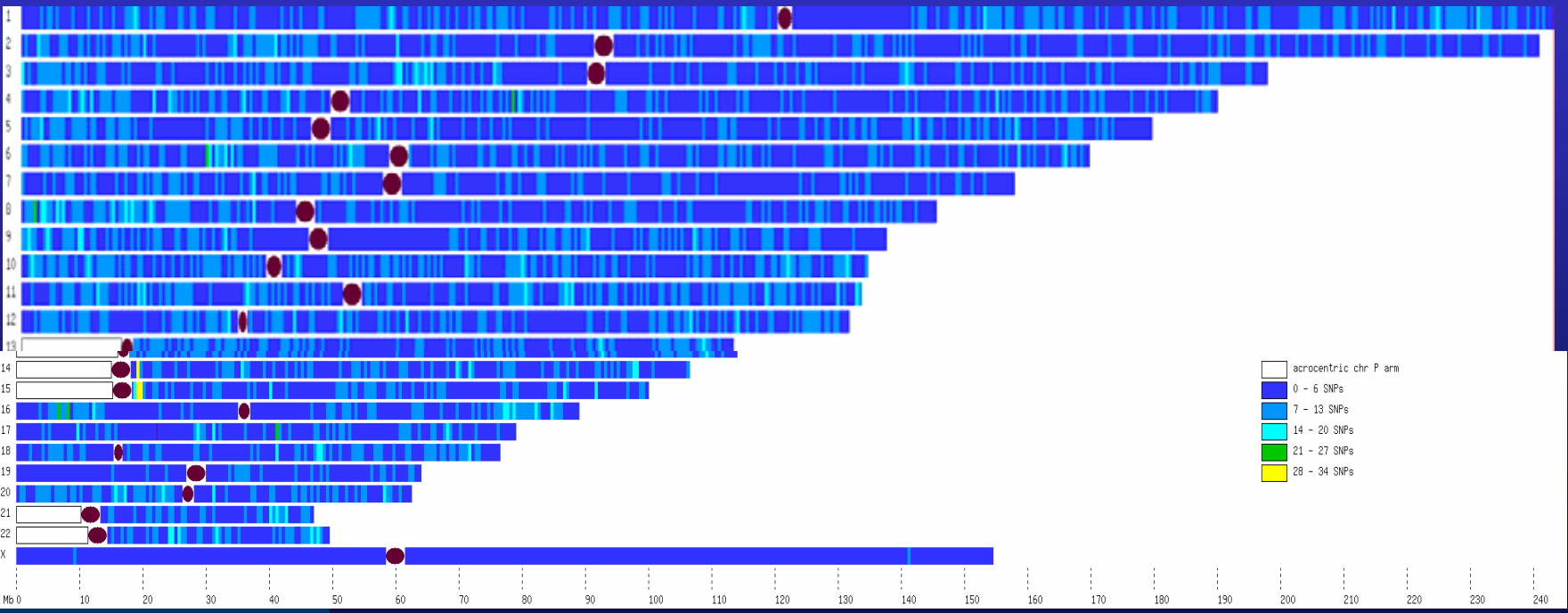
Vol 447|7 June 2007|doi:10.1038/nature05911 nature

ARTICLES

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

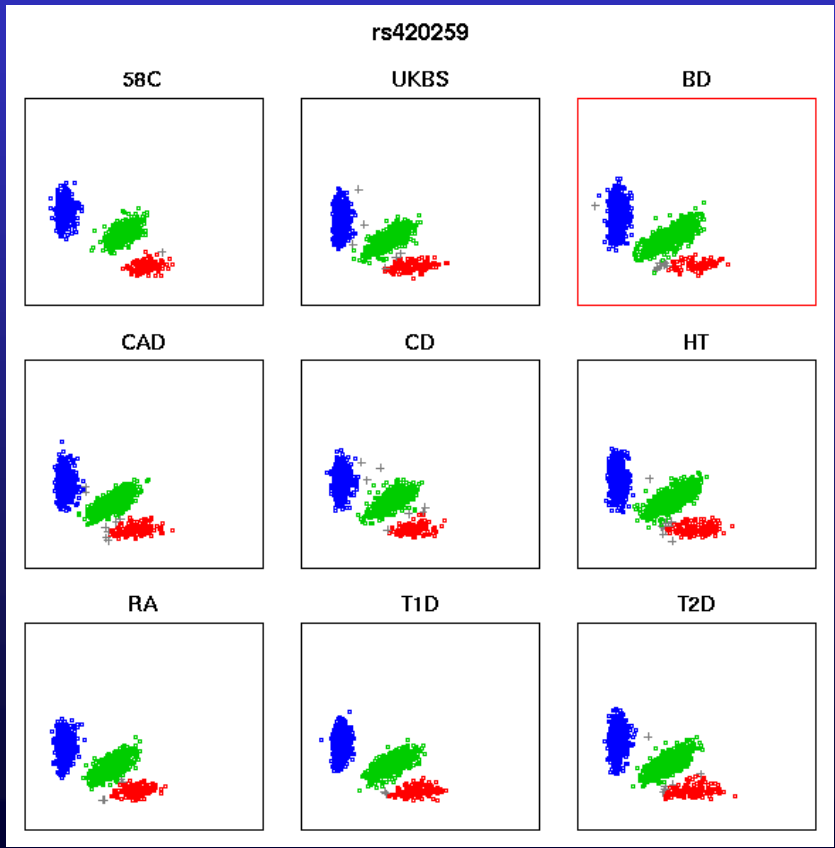
The Wellcome Trust Case Control Consortium*

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 $P < 5 \times 10^{-7}$: 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^{-5} and 5×10^{-7}) likely to yield



391,000 SNPs minor allele frequency > 1% and passing quality control, Affymetrix 500K SNP Chip

- 14,000 cases, 3,000 controls
- 809 excluded based on non Caucasian ancestry, contamination, false identity, relatedness


rs420259
58C
UKBS
BD
CAD
CD
HT
RA
T1D
T2D

Their results

- 11 of 12 known disease SNPs showed effects
- At $p < 5 \times 10^{-7}$ cut-off, results for all SNPs analysed:

disease	sibling risk	significant SNPs
Crohns disease	> 17	9
T1 Diabetes	15	7
T2 Diabetes	3	3
Rheumatoid arthritis	> 5	3
bipolar disorder	> 7	1
coronary artery disease	> 2.5	1
hypertension	> 2.5	0

Sample size:
3,000 controls
2,000 of each disease

12 of the SNPs that showed strong geographic variation

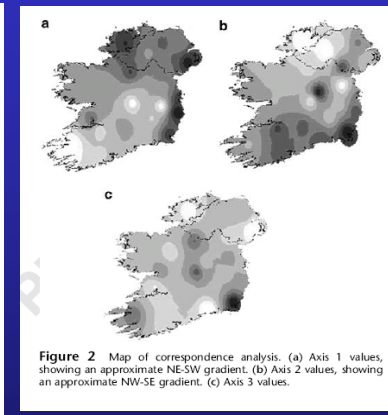


Figure 2 Map of correspondence analysis. (a) Axis 1 values, showing an approximate NE-SW gradient. (b) Axis 2 values, showing an approximate NW-SE gradient. (c) Axis 3 values.

Ciara Dolan et al
Eur J Hum Genet
2005

Nature 2007 447:661 Wellcome Trust Case Control Consortium

Genotyping polymorphisms

- There are many methods
- You can outsource this or do it in-house
- For a typical single nucleotide polymorphism there are two variants, and three genotypes
  - E.g. at position 1 of codon 1205 of COL3A1 gene
    - more common allele G encoding Valine
    - rarer allele A which results in change to Isoleucine
      - Three potential genotypes
        - » GG
        - » GA
        - » AA

So I know which SNPs I want to type, how do I genotype them?

- We used to use K-Biosciences
  - Post DNA
  - E-mail dbSNP rs number and sequence
  - Results returned by e-mail
    - *If* it works, cheaper than doing it yourself.
    - Very low consumption of DNA (4 ng per assay)
      - Miniaturisation (reagent costs go down too)
- Illumina SNP genotyping platform
  - Local or outsource to the company
- Many other methods:
    - e.g restriction enzymes and gel sizing
- Direct Sequencing now cost effective for certain study designs

Association Analysis of genotypes

- Compare frequency of A allele in case and controls
- Compare three genotype frequencies in cases and controls
    - e.g. chi-square with 2 degrees of freedom
- Investigate frequency of one genotype (e.g. AA versus AG+GG)
  - If allele A is dominant
- Trend test: assumes an allele dosage effect
  - one degree of freedom.
- Usual statistical measures: Odds Ratio or Relative Risk, with 95% confidence intervals.

SNP association studies

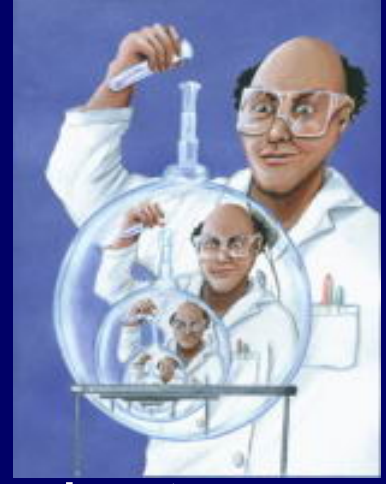
- Very frequently it is not the variant that you type that causes the difference
- Instead, a variant in association with the typed variant may cause the effect

Sometimes, want to analyse whether a particular combination typed of variants at a locus has an effect (typically the COMBINATION is associated strongly with a rarer variant that is not typed).

Association Studies analysing more than one genotype

- Haplotype analysis
  - Genotypes of variants which are chromosomally very close (e.g. within 100 kilobases)
    - Close variants in “linkage disequilibrium”, co-occur within individuals in population more than expected
      - » reflects accidents of mutation, recombination, selection & genetic drift during history of the human population
  - Specialised analytical methods
    - e.g. analyse the most frequent haplotypes
    - Get a statistical programme to both:
      - » estimate what are the most common haplotypes
      - » judge if these are associated with disease

Power



- The bigger the better
  - E.g. 2,000 patients, 2,000 controls for common diseases
- Large-throughput screens of many variants
  - findings will require replication/meta-analysis
    - Meta analysis: putting together the association results of many studies, and testing significance

Study design

Are assumptions correct?

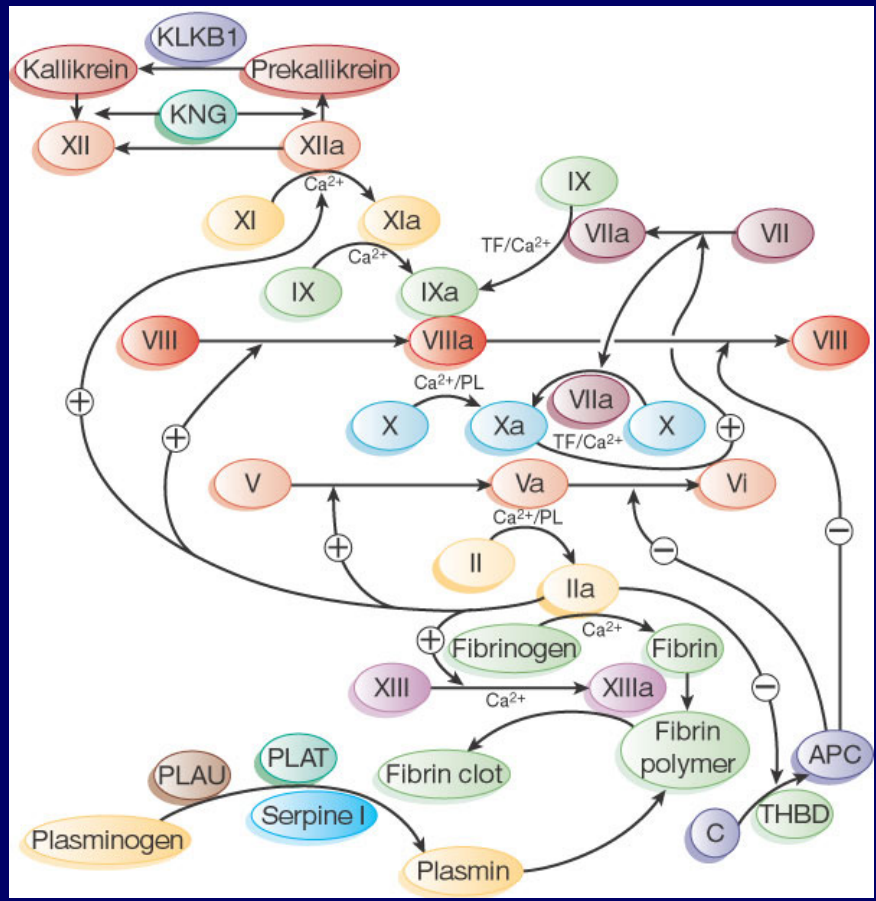
- Will common haplotypes defined by hapmap project capture the variation you need to study?
- Is this candidate gene really that interesting?
- Are effects of common SNPs in this disease
  - » Too small to ever say anything worth studying?
- Can you a priori identify the subset of SNPs with bigger biological effects?
  - » Polyphen database: provides a guess of whether amino-acid altering SNPs are benign or damaging
  - » www.bork.embl-heidelberg.de/PolyPhen/

Study analysis

- Need to think carefully about analysis
  - What programmes do is essentially not too complex
  - However, it is very easy to cheat to get a significant p-value that is meaningless.
  - Need to define analysis plan in advance, rather than find the analysis that gives you the result you want.

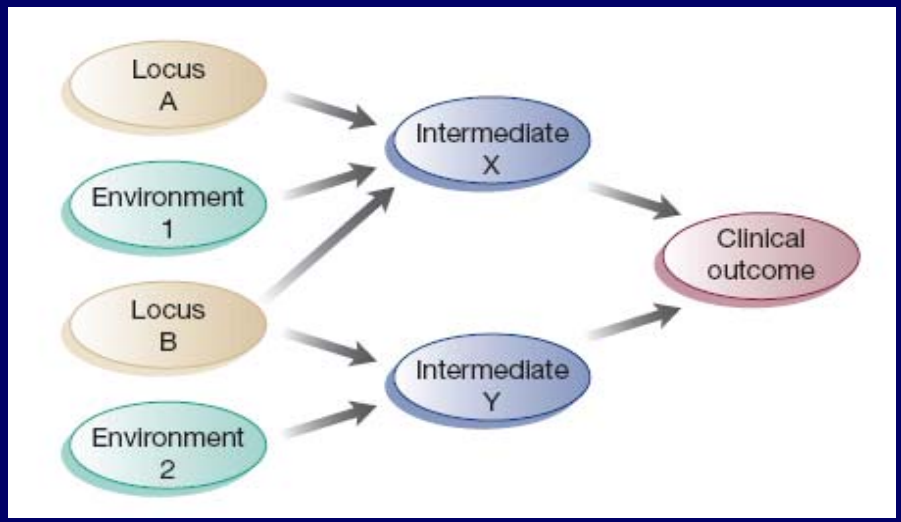
Association Studies: interactions

- Gene-gene interaction
  - Typically, genes whose interaction we are interested in are not close enough to be in disequilibrium
  - Very large samples needed to detect significant interactions



- Gene environment interaction
  - Still need bigger samples
  - E.g. clinical trial power is chosen
    - to be enough for main effect of drug on outcome
      - not for testing interaction with genes etc

Using genetics to tease apart Intermediate phenotypes



- POSSIBLE SCENARIO:
- (1) You have a risk factor associated with disease
  - E.g. fibrinogen and heart disease
- (2) A genetic variant is associated with risk factor
  - E.g. fibrinogen promoter variant
- (3) The genetic variant is not associated with disease
  - E.g. fibrinogen promoter variant and heart disease
- 1+2+3 = CONCLUDE:
  - fibrinogen is caused by, and does not cause disease
- Under a simple model, this is true
- This implies that negative genetic associations with common disease can tell you something.
- *For discussion of this "Mendelian randomisation", see: Lancet 2003;9388:930-931*

Resources to think about

- Affymetrix 10k SNP chip
  - Amino acid altering
- Affymetrix 1M 500K SNP chip
- Illumina 1M SNP chip
- Gene copy number variation assays
  - SNP chip or specialised
- Other platforms also: prices going down all the time
- Is your study for
  - Primary discovery
  - Replication of findings from largescale studies
  - Replication of finding from smallscale studies
  - Phenotype well studied elsewhere
  - Unusual, hard to replicate phenotype

Sample of Dublin people working in this area

- Genetic epidemiology (statistical analysis & study design)
  - Gianpiero Cavalleri (RCSI)
  - Me
- Lab aspects
  - Derek Morris (TCD, Michael Gill's group)
  - Ross McManus (TCD)
  - Sean Ennis (UCD) Illumuna platform
- Clinical experience, association studies
  - Michael Gill, TCD: schizophrenia/bipolar etc
  - Norman Delanty/Colin Doherty RCSI/TCD epilepsy
  - Alice Stanton RCSI, Helen Colhoun UCD (clinical trial association)
- Linkage analysis
  - David Barton, Crumlin UCD
  - Pete Humphries group TCD

Microarray refinement of genes within linkage regions