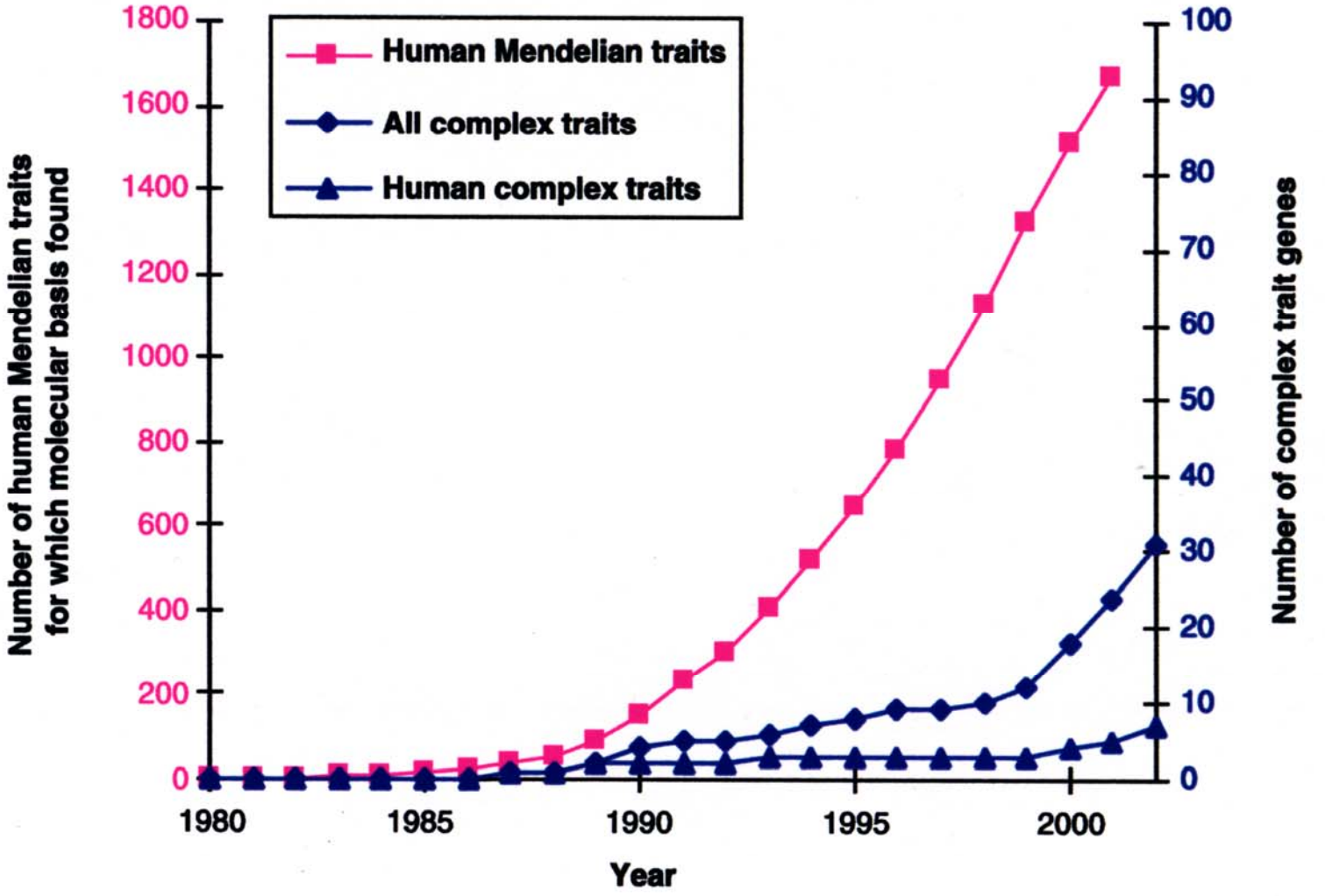
Primer on Genome Wide Association and Sequencing

Multi-IC NIH Symposium On Population Genomics

Francis S. Collins, M.D., Ph.D.

June 5, 2006



Glazier et al., Science 298:2345-9, 2002

Association is much more powerful than linkage to identify common susceptibility variants

N. Risch and K. Merikangas
Science 273: 1516-1517, 1996

(As long as you're looking for common alleles)

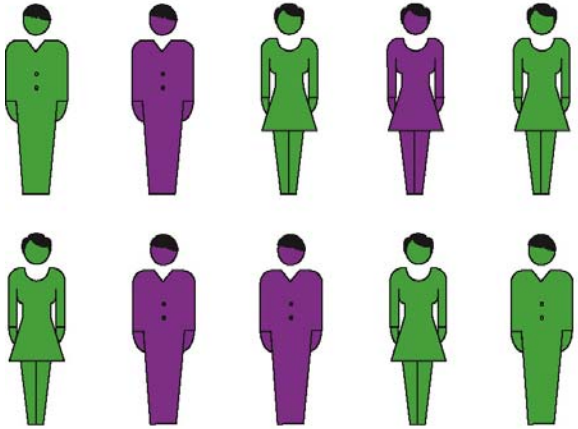
But until recently, association studies have only been practical when a candidate gene was suspected

-- and usually we're not smart enough to pick the right candidates

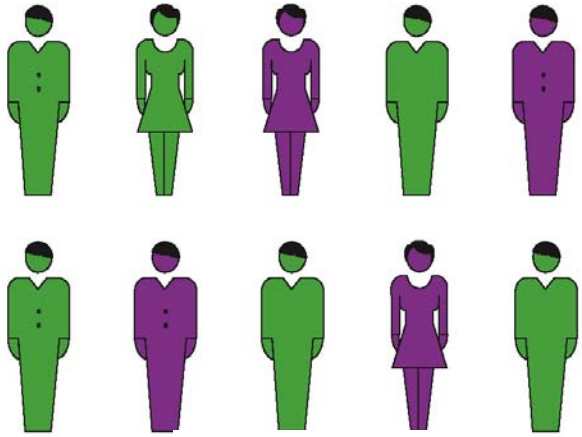
Sequence from chromosome 7

Three single nucleotide polymorphisms (SNPs) are present

SNP A

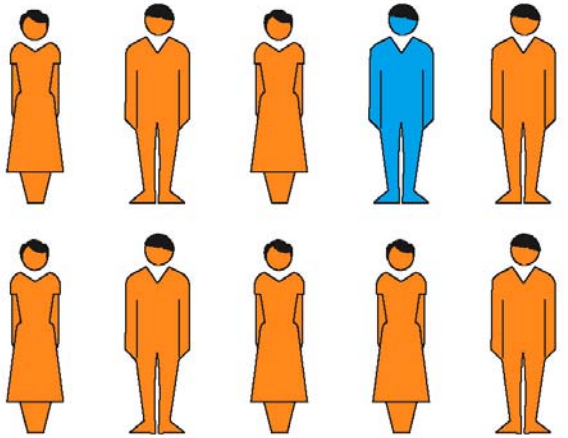


Diabetic

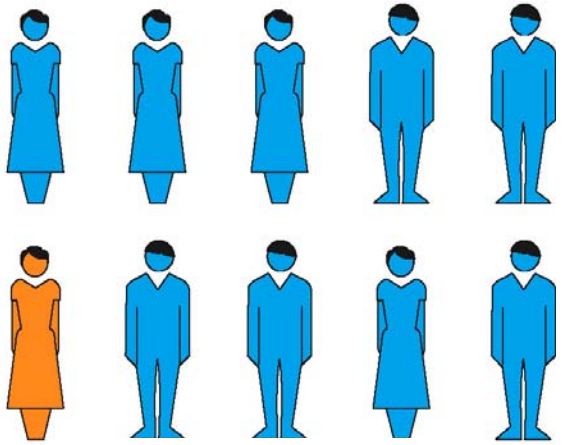


Unaffected

SNP B



Diabetic



Unaffected

"Whole Genome Association" Approach to Common Disease: The View from 2002

- Identify all 10 million common SNPs
- Collect 1000 cases and 1000 controls
- Genotype all DNAs for all SNPs
- That adds up to 20 billion genotypes
- At 50 cents a genotype, that's $10 billion for each disease – completely out of the question

Sequence from chromosome 7

GAAATAATTAATGTTTTCCTTCCTTCTCCTATTTTGTCCTTTACTTCAATTTATTTATTTATTATTAATATTATTATTTTTTG
AGACGGAGTTTCACTCTTGTTGCCAACCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCACACTCCGCTTTC**C/T**GG
TTTCAAGCGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGACTACAGTCACACACCACCACGCCCGGCTAATTTTTG
TATTTTTAGTAGAGTTGGGGTTTCACCATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCCGCCAGCCTCT
GCCTCCCAAAGAGCTGGGATTACAGGCGTGAGCCACCGCGCTCGGCCCTTTGCATCAATTTCTACAGCTTGTTTTCTT
TGCCTGGACTTTACAAGTCTTACCTTGTTCTGCCTTCAGATATTTGTGTGGTCTCATTCTGCTTGTGCCAGTAGCTAAAA
ATCCATGATTTGCTCTCATCCCACTCCTGTTGTTCATCTCCTCTTATCTGGGGTCAC**A/C**TATCTCTTCGTGATTGCATTC
TGATCCCCAGTACTTAGCATGTGCGTAACAACTCTGCCTCTGCTTTCCCAGGCTGTTGATGGGGTGCTGTTCATGCCT
CAGAAAAATGCATTGTAAGTTAAATTATTAAAGATTTTAAATATAGGAAAAAAGTAAGCAAACATAAGGAACAAAAAG
GAAAGAACATGTATTCTAATCCATTATTTATTATACAATTAAGAAATTTGGAAACTTTAGATTACACTGCTTTTAGAGAT
GGAGATGTAGTAAGTCTTTTACTCTTTACAAAATACATGTGTTAGCAATTTTGGGAAGAATAGTAACTCACCCGAACA
GTGTAATGTGAATATGTCACTTACTAGAGGAAAGAAGGCACTTGAAAAACATCTCTAAACCGTATAAAAACAATTACA
TCATAATGATGAAAACCCAAGGAATTTTTTTAGAAAACATTACCAGGGCTAATAACAAAGTAGAGCCACATGTCATTT
ATCTTCCCTTTGTGTCTGTGTGAGAATTCTAGAGTTATATTTGTACATAGCATGGAAAAATGAGAGGCTAGTTTATCAA
CTAGTTCATTTTTAAAAGTCTAACACATCCTAGGTATAGGTGAACTGTCCTCCTGCCAATGTATTGCACATTTGTGCCC
AGATCCAGCATAGGGTATGTTTGCCATTTACAAACGTTTATGTCTTAAGAGAGGAAATATGAAGAGCAAAACAGTGCA
TGCTGGAGAGAGAAAGCTGATACAAATATAAATGAAACAATAATTGGAAAAATTGAGAAACTACTCATTTTCTAAATT
ACTCATGTATTTTCCTAGAATTTAAGTCTTTTAATTTTTGATAAATCCCAATGTGAGACAAGATAAGTATTAGTGATGGT
ATGAGTAATTAATATCTGTTATATAATATTCATTTTCATAGTGGAAGAAATAAAATAAAGGTTGTGATGATTGTTGATTA
TTTTTTCTAGAGGGGTTGTCAGGGAAAGAAATTGCTTTTTTTCATTCTCTCTTTCCACTAAGAAAGTTCAACTATTAATT
TAGGCACATACAATAATTACTCCATTCTAAAATGCCAAAAAGGTAATTTAAGAGACTTAAAACTGAAAAGTTTAAGATA
GTCACACTGAACTATATTAAAAAATCCACAGGGTGGTTGGAACTAGGCCTTATATTAAAGAGGCTAAAAATTGCAATA
AGACCACAGGCTTTAAATATGGCTTTAAACTGTGAAAGGTGAAACTAGAATGAATAAAATCCTATAAATTTAAATCAA
AAGAAAGAAACAAACT**A/G**AAATTAAAGTTAATATACAAGAATATGGTGGCCTGGATCTAGTGAACATATAGTAAAGA
TAAAACAGAATATTTCTGAAAAATCCTGGAAAATCTTTTGGGCTAACCTGAAAACAGTATATTTGAAACTATTTTTAAA

Are the SNPs correlated with their neighbors?

These three SNPs could theoretically occur in 8 different haplotypes

…C…A…A…

…C…A…G…

…C…C…A…

…C…C…G…

…T…A…A…

…T…A…G…

…T…C…A…

…T…C…G…

But in practice, only two are observed

…C…A…A…

…C…A…G…

…C…C…A…

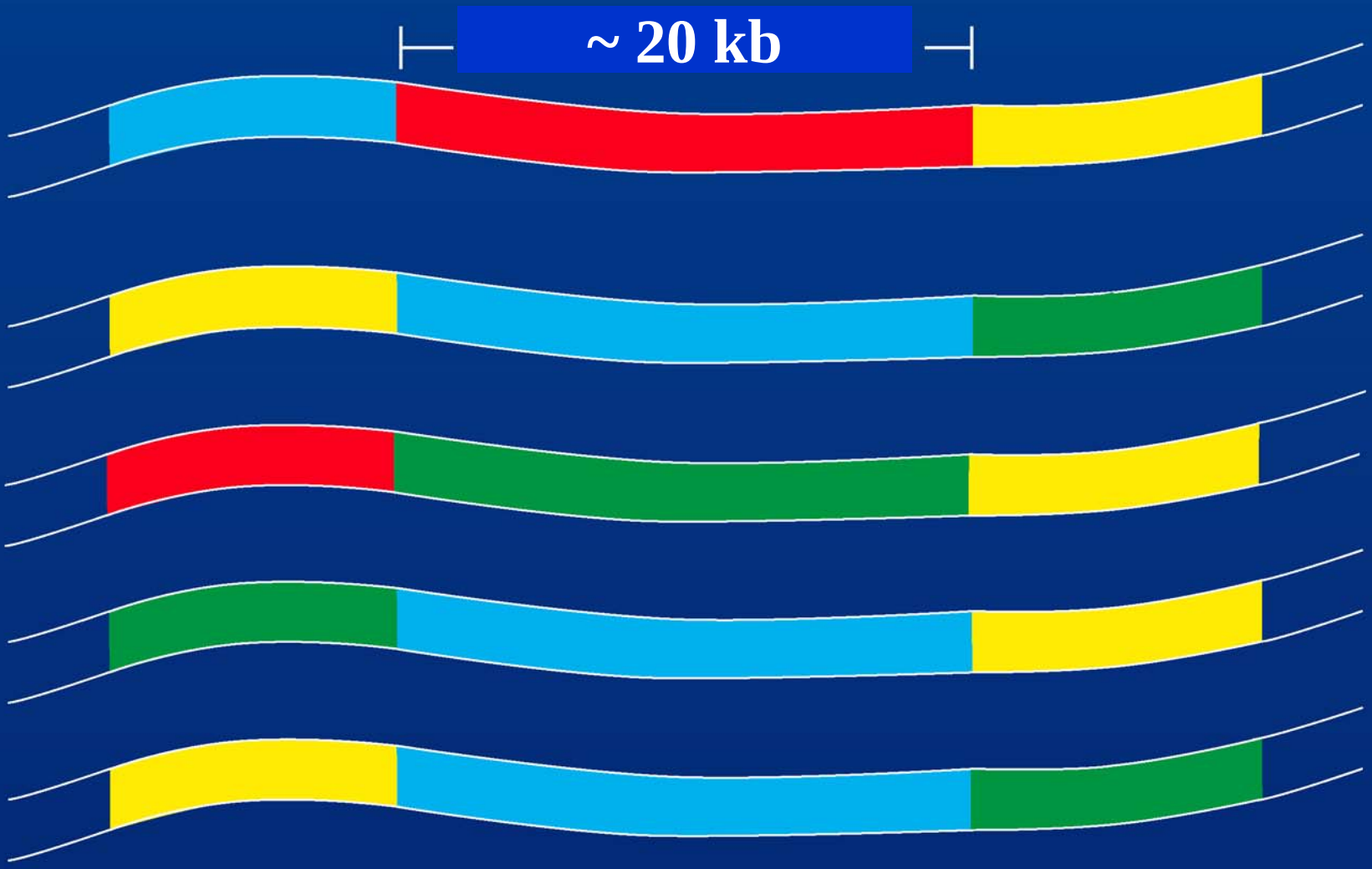
…C…C…G…

…T…A…A…

…T…A…G…

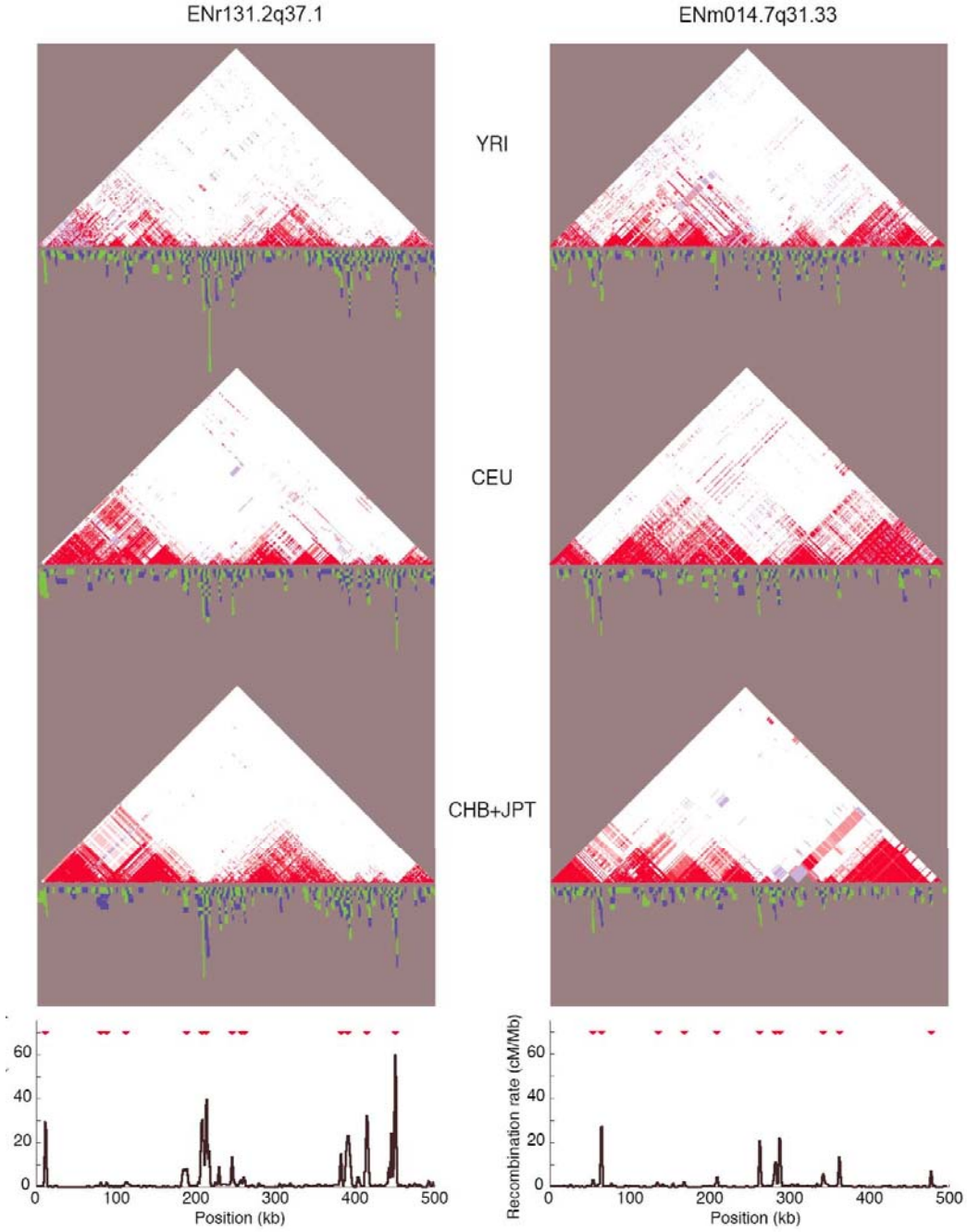
…T…C…A…

…T…C…G…


~ 20 kb

International HapMap Consortium:

Canada, China, Japan, Nigeria,
United Kingdom, United States


ENr131.2q37.1
ENm014.7q31.33
YRI
CEU
CHB+JPT
Recombination rate (cM/Mb)
Position (kb)
Position (kb)

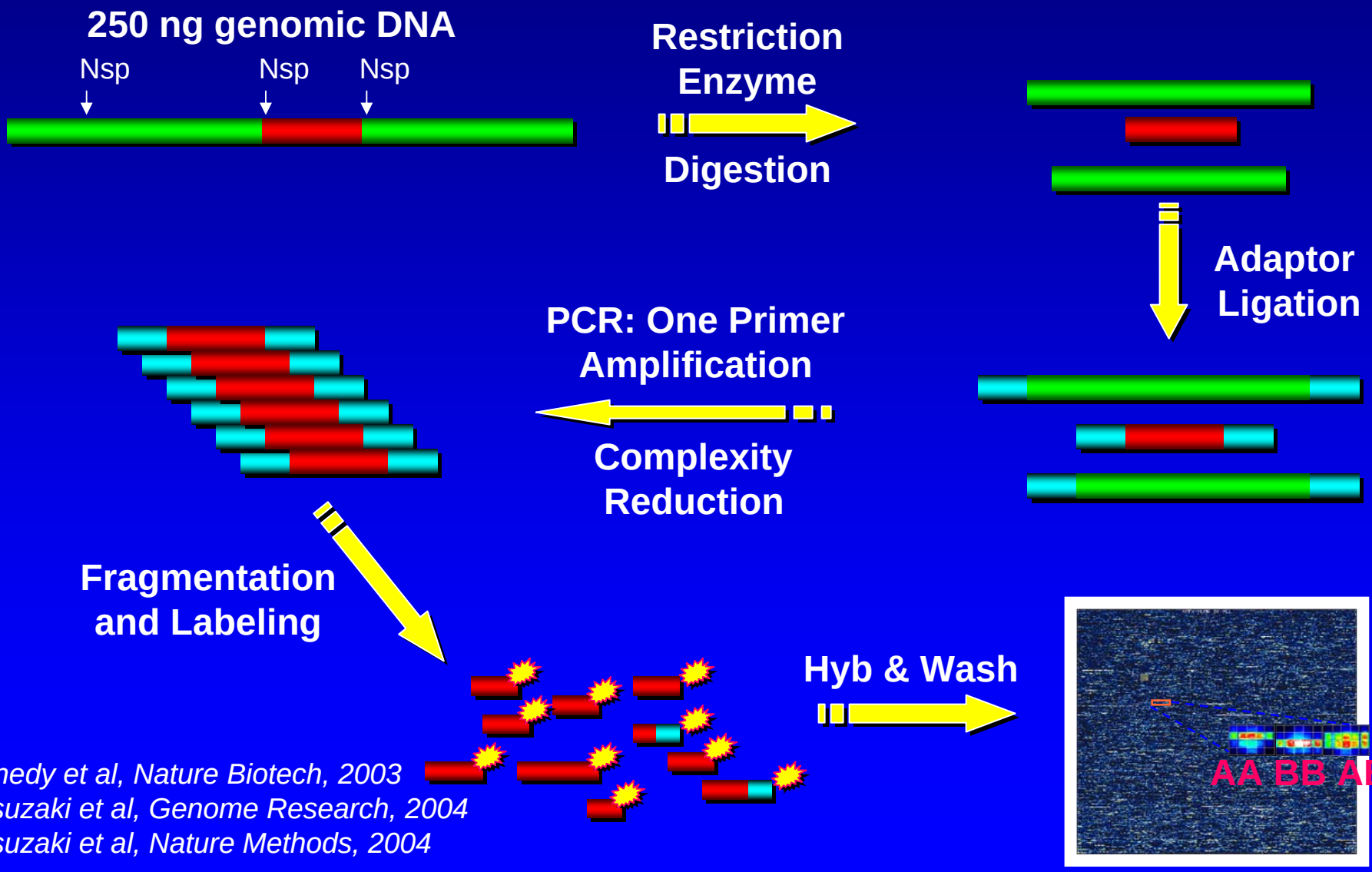
"Whole Genome Association" Approach to Common Disease in the HapMap Era

- Identify an optimum set of 300,000 tag SNPs
- Collect 1000 cases and 1000 controls
- Genotype all DNAs for all SNPs
- That adds up to 600 million genotypes
- This would still be too expensive if a genotype cost $0.50, but there have been other developments…..

Costs of Large Scale SNP Genotyping Have Come Down by More Than 100-fold in 4 Years

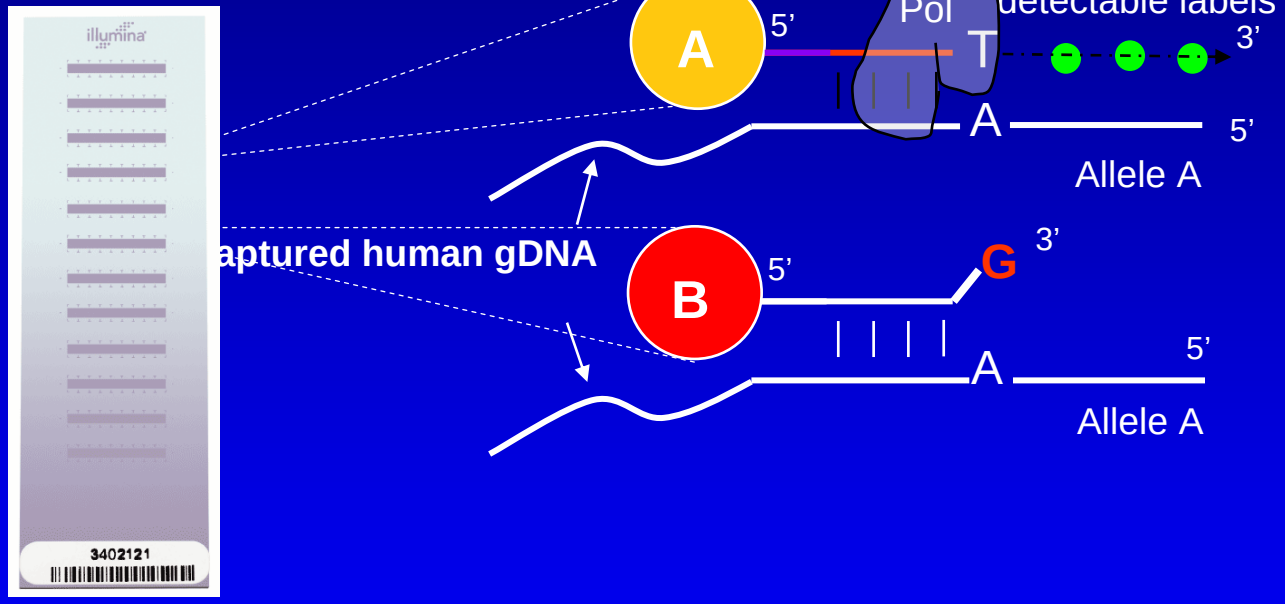
- Affymetrix 100K -> 500K
- Illumina 100K -> 317K -> 550K
- Perlegen (in house only)

Affymetrix GeneChip® Mapping Assay Overview

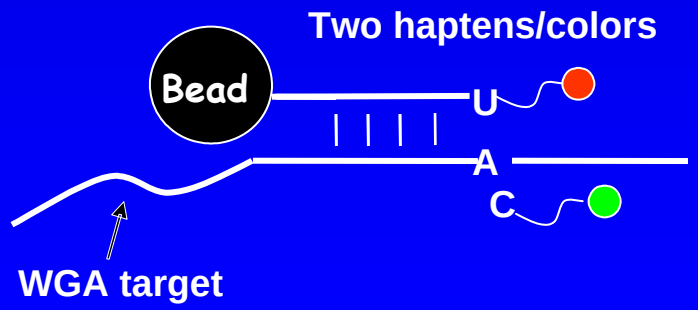


Kennedy et al, Nature Biotech, 2003
Matsuzaki et al, Genome Research, 2004
Matsuzaki et al, Nature Methods, 2004

Illumina Infinium Chemistry



Infinium I:
Allele Specific
Extension



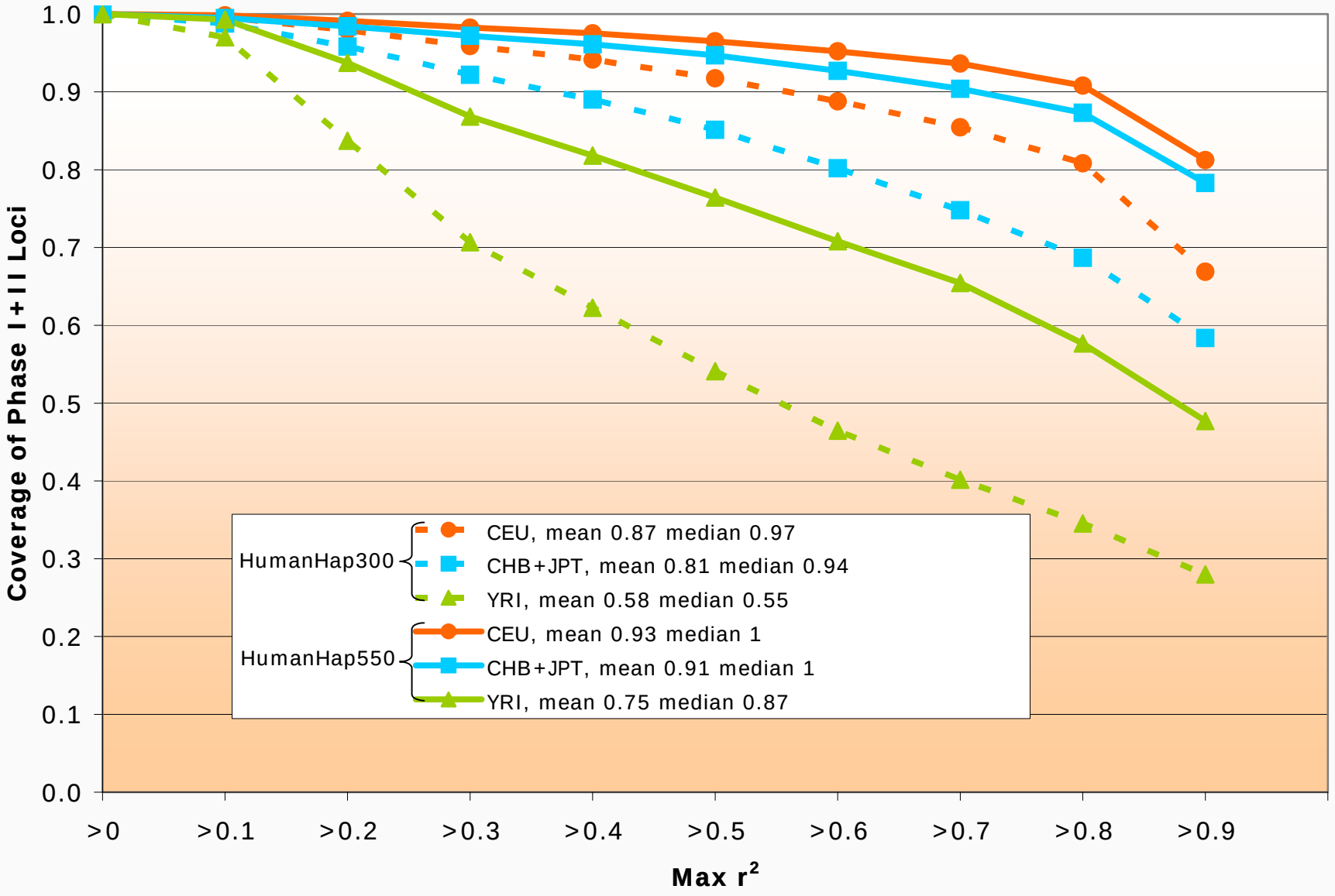
Infinium II:
Single Base
Extension

AREDS 100K Quality Comparisons (CIDR)

	Affymetrix 100K BRLMM	Illumina Human-1
# loci attempted	**116,190**	**109,365**
# chips/sample	**2**	**1**
# samples attempted	**633**	**630**
# samples dropped	**87**	**7**
Redo required	**14%**	**5.2%**
Call rate	**99.5%**	**99.8%**
Reproducibility	**99.85%**	**99.997%**
Mendelian consistency rate (trios)	**99.54%**	**99.995%**
HapMap concordance	**99.75%**	**99.64%**

Provided by Kim Doheny, CIDR

Coverage of Phase I and Phase II Hapmap loci with MAF > 0.05: Illumina HumanHap300 and HumanHap550 by Population



Provided by Sarah Murray, Illumina

Other Important Issues In Designing An Association Study

- **Quality of phenotypes**

Other Important Issues In Designing An Association Study

- Quality of phenotypes
- **Power estimates**

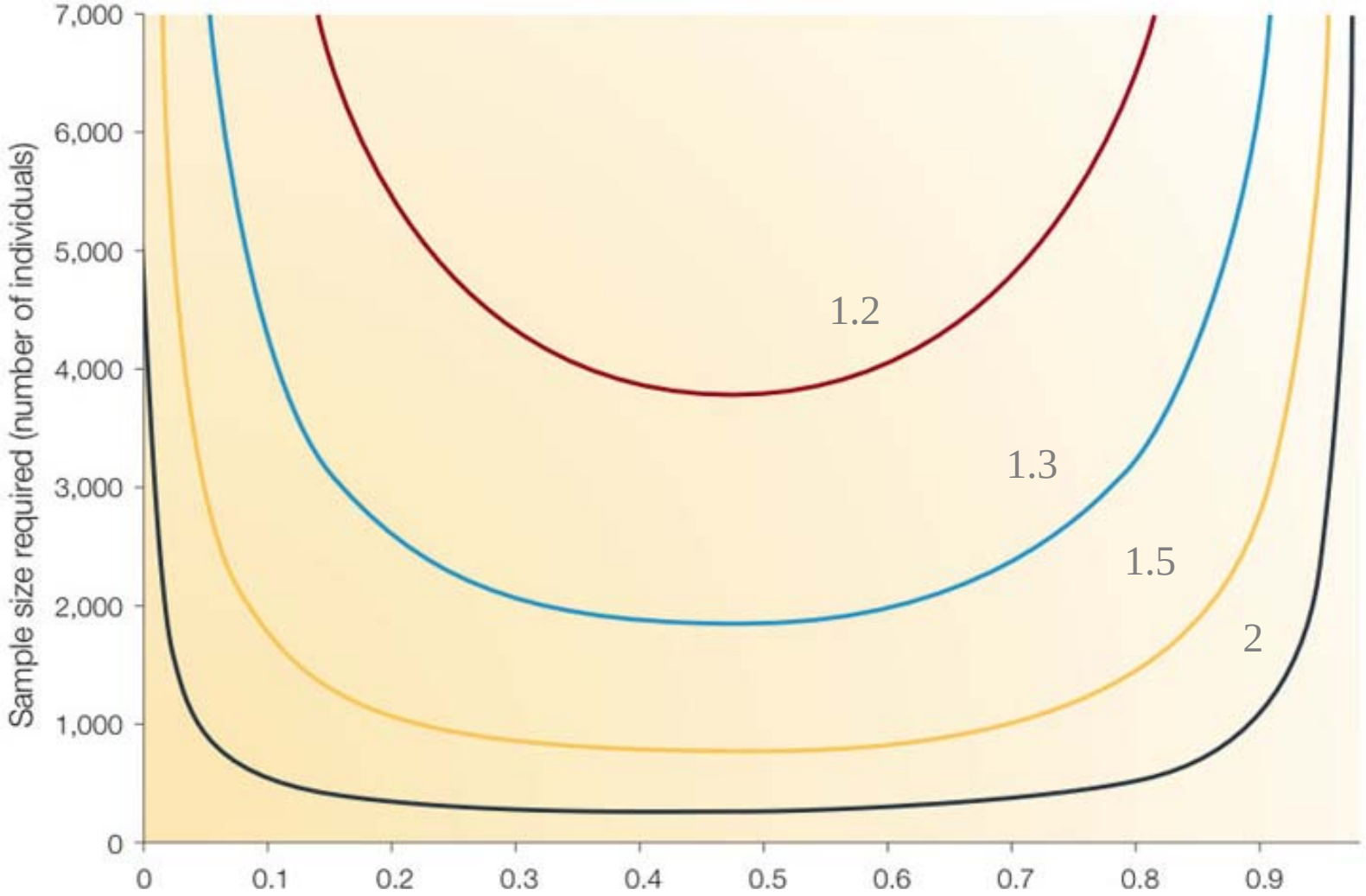


Figure 1 | **Effects of allele frequency on sample-size requirements.** The numbers of cases and controls that are required in an association study to detect disease variants with allelic odds ratios of 1.2 (red), 1.3 (blue), 1.5 (yellow) and 2 (black) are shown. Numbers shown are for a statistical power of 80% at a significance level of $P < 10^{-6}$, assuming a multiplicative model for the effects of alleles and perfect correlative linkage disequlibrium between alleles of test markers and disease variants.

Source: Wang et al., Nature Reviews Genetics, 2005

Other Important Issues In Designing An Association Study

- Quality of phenotypes
- Power estimates
- **Staged design**

nature genetics

Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies

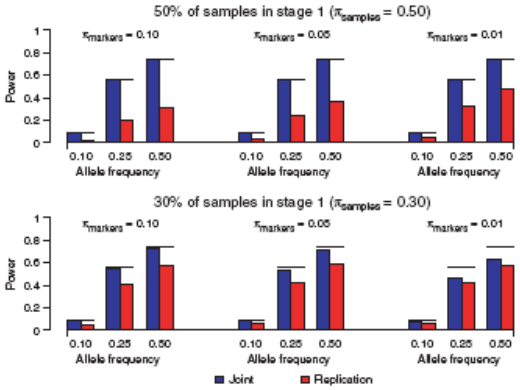
Andrew D Skol, Laura J Scott, Gonçalo R Abecasis & Michael Boehnke

Genome-wide association is a promising approach to identify common genetic variants that predispose to human disease[1–4]. Because of the high cost of genotyping hundreds of thousands of markers on thousands of subjects, genome-wide association studies often follow a staged design in which a proportion ($\pi_{samples}$) of the available samples are genotyped on a large number of markers in stage 1, and a proportion ($\pi_{samples}$) of these markers are later followed up by genotyping them on the remaining samples in stage 2. The standard strategy for analyzing such two-stage data is to view stage 2 as a replication study and focus on findings that reach statistical significance when stage 2 data are considered alone[2]. We demonstrate that the alternative strategy of jointly analyzing the data from both stages almost always results in increased power to detect genetic association, despite the need to use more stringent significance levels, even when effect sizes differ between the two stages. We recommend joint analysis for all two-stage genome-wide association studies, especially when a relatively large proportion of the samples are genotyped in stage 1 ($\pi_{samples} \geq 0.30$), and a relatively large proportion of markers are selected for follow-up in stage 2 ($\pi_{markers} \geq 0.01$).

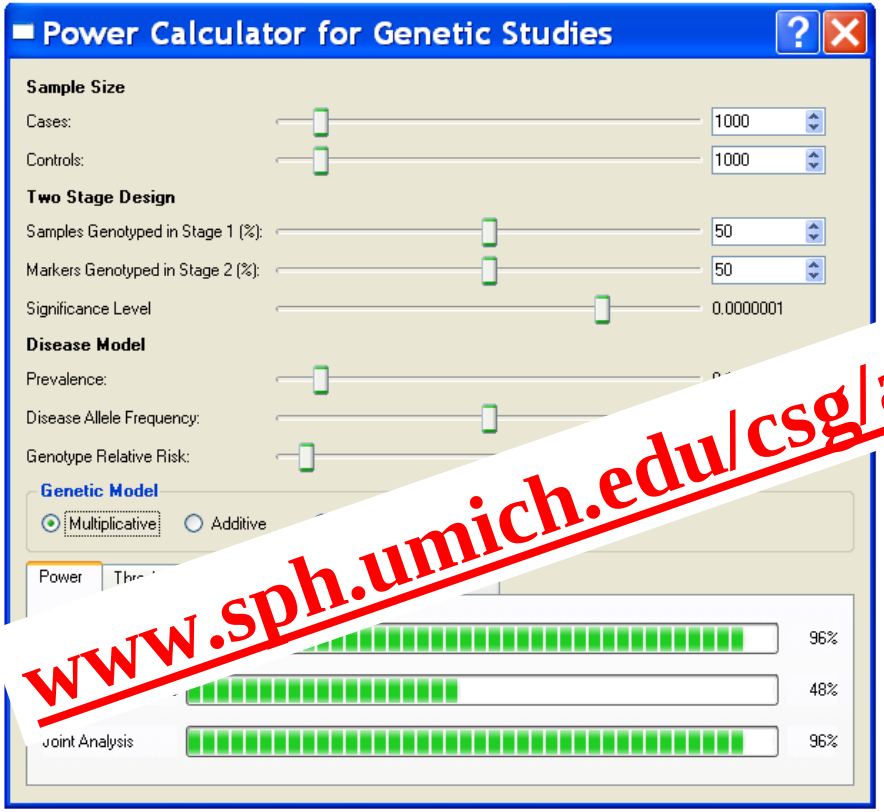
Genome-wide association studies are now underway[5], enabled by rapidly decreasing genotyping costs, massively multiplexed genotyping technologies and the large-scale SNP discovery and genotyping efforts of the SNP Consortium[6], the HapMap project[7] and Perlegen Sciences[8]. These projects have identified and genotyped well over 1 million SNPs in several human populations, allowing investigators to select a set of genetic markers that efficiently assays most common human genetic variation[9–11]. Compared with one-stage designs that genotype all samples on all markers, well-constructed two-stage association designs maintain power while substantially reducing

We focus on two-stage designs in which all M markers are genotyped in a proportion of the samples ($\pi_{samples}$) in stage 1, and results of stage 1 are used to select a proportion of these M markers ($\pi_{markers}$) for follow-up on the remaining samples in stage 2. These samples might be cases and controls for a genetic disease or individuals measured for a quantitative trait. We assume initially that the M markers are in linkage equilibrium.

Our purpose is to compare power for the standard replication-based analysis strategy with the power of the alternative strategy of joint analysis of all available samples. Both strategies can be tailored to achieve any desired genome-wide false positive rate (type I error rate) of α_{genome} so that the number of false positives expected in the genome-wide association scan is α_{genome}. In the replication strategy,



CaTS Tutorial



Brief Introduction

CaTS is an interactive power calculator for genetic association studies, with features that facilitate the design of two-stage genetic association studies.

Use the sliders and input boxes in the top half of the CaTS dialog to setup your study design and disease model parameters, and then select one of the tabs at the bottom of the screen to get estimated power, recommended thresholds for two stage analysis, and the fitted penetrances for each genotype.

Other Important Issues In Designing An Association Study

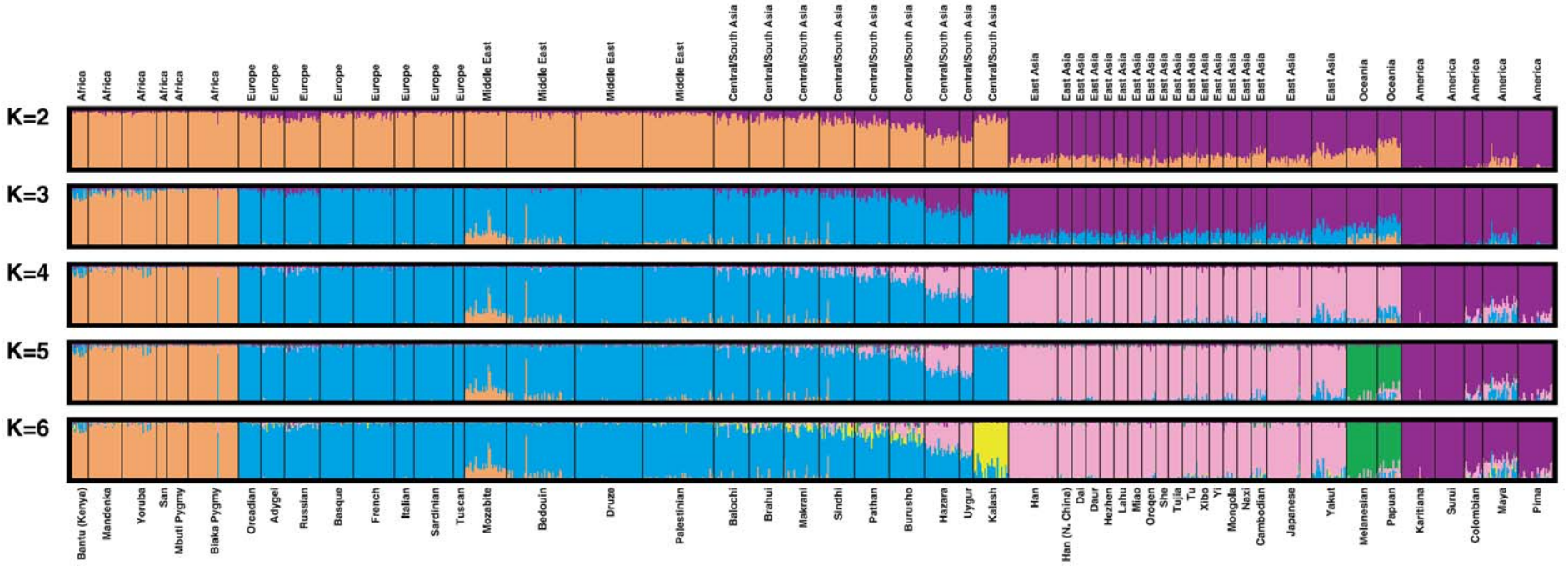
- Quality of phenotypes
- Power estimates
- Staged design
- Possibility of stratification/mismatching

Genetic Structure of Human Populations

Noah A. Rosenberg,[1]* Jonathan K. Pritchard,[2] James L. Weber,[3] Howard M. Cann,[4] Kenneth K. Kidd,[5] Lev A. Zhivotovsky,[6] Marcus W. Feldman[7]

We studied human population structure using genotypes at 377 autosomal microsatellite loci in 1056 individuals from 52 populations. Within-population differences among individuals account for 93 to 95% of genetic variation; differences among major groups constitute only 3 to 5%. Nevertheless, without using prior information about the origins of individuals, we identified six main genetic clusters, five of which correspond to major geographic regions, and subclusters that often correspond to individual populations. General agreement of genetic and predefined populations suggests that self-reported ancestry can facilitate assessments of epidemiological risks but does not obviate the need to use genetic information in genetic association studies.

www.sciencemag.org SCIENCE VOL 298 20 DECEMBER 2002



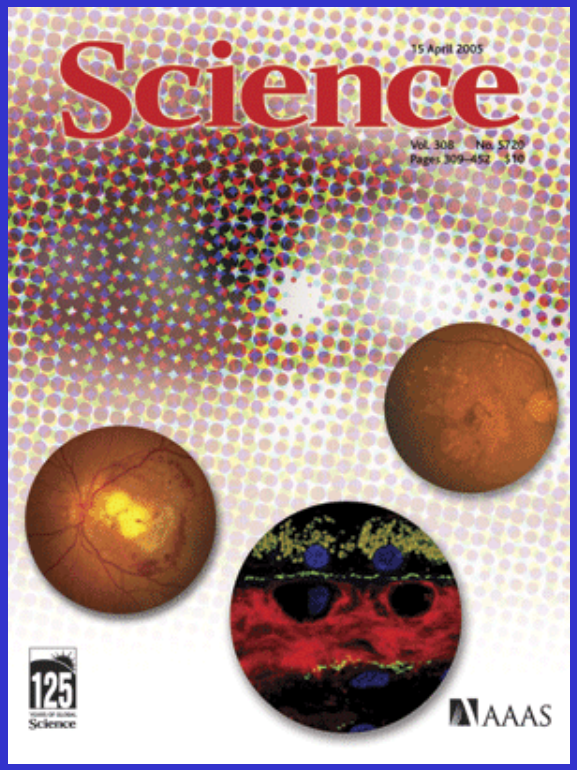
Other Important Issues In Designing An Association Study

- Quality of phenotypes
- Power estimates
- Staged design
- Possibility of stratification/mismatching
- **Case-control vs. family-based**

The First HapMap Success Story: Age-Related Macular Degeneration

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7†]



Two other risk variants have now been identified.

Together these account for 74% of risk, and point to powerful new approaches to prevention and treatment.

nature genetics

HapMap leads to a new diabetes gene discovery

Variant of transcription factor 7-like 2 (*TCF7L2*) confers risk of type 2 diabetes

Struan F A Grant[1], Gudmar Thorleifsson[1], Inga Reynisdottir[1], Rafn Benediktsson[2,3], Andrei Manolescu[1], Jesus Sainz[1], Agnar Helgason[1], Hreinn Stefansson[1], Valur Emilsson[1], Anna Helgadottir[1], Unnur Styrkarsdottir[1], Kristinn P Magnusson[1], G Bragi Walters[1], Ebba Palsdottir[1], Thorbjorg Jonsdottir[1], Thorunn Gudmundsdottir[1], Arnaldur Gylfason[1], Jona Saemundsdottir[1], Robert L Wilensky[4], Muredach P Reilly[4], Daniel J Rader[4], Yu Bagger[5], Claus Christiansen[5], Vilmundur Gudnason[2], Gunnar Sigurdsson[2,3], Unnur Thorsteinsdottir[1], Jeffrey R Gulcher[1], Augustine Kong[1] & Kari Stefansson[1]

This result has been already confirmed by multiple groups in diverse populations

Association of DG8S737 "Allele – 8" with Prostate Cancer

	Allele Frequency			
	Cases	Controls	OR	P-value
Iceland	0.131	0.078	1.77	$2x10^{-8}$
Sweden	0.101	0.079	1.38	$4x10^{-3}$
Chicago (European)	0.082	0.041	2.10	$3x10^{-3}$
Michigan (African-American)	0.234	0.161	1.60	$2x10^{-3}$

Amundadottir et al., Nature Genetics 38:652-8, 2006

	Single common allele	Multiple rare alleles
Mendelian		
Polygenic		

	Single common allele	Multiple rare alleles
Mendelian		
Polygenic		

LINKAGE

	Single common allele	Multiple rare alleles
Mendelian		
Polygenic		

LINKAGE

	Single common allele	Multiple rare alleles
Mendelian		
Polygenic		

ASSOCIATION

	Single common allele	Multiple rare alleles
Mendelian		
Polygenic		

ASSOCIATION

	Single common allele	Multiple rare alleles
Mendelian		
Polygenic		

SEQUENCING

	Single common allele	Multiple rare alleles
Mendelian		
Polygenic		

SEQUENCING

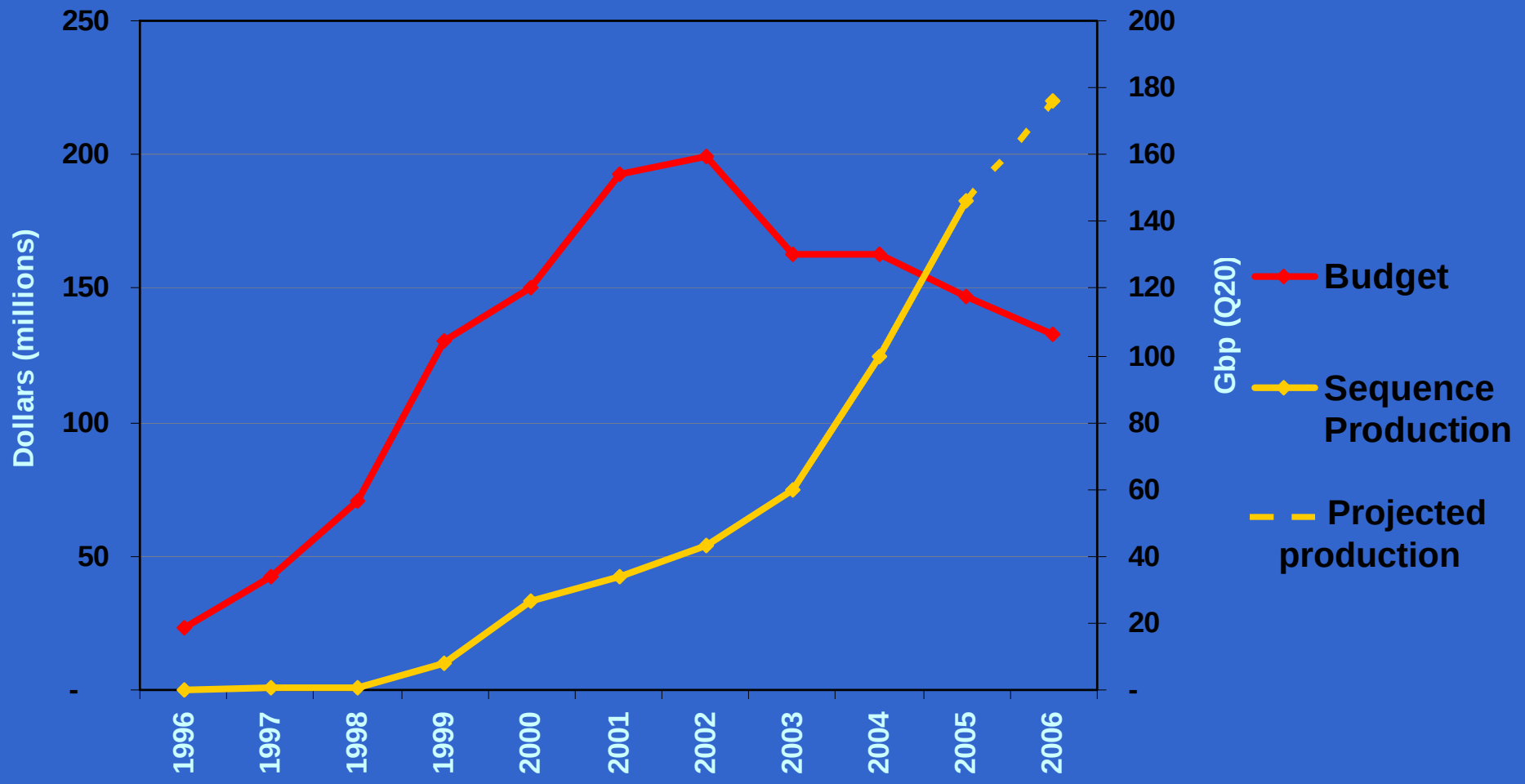
Multiple Rare Alleles Contribute to Low Plasma Levels of HDL Cholesterol

Jonathan C. Cohen,[1,2,3*†] Robert S. Kiss,[5*] Alexander Pertsemlidis,[1] Yves L. Marcel,[5†] Ruth McPherson,[5] Helen H. Hobbs[1,3,4]

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.

Science 305: 869-872, 2004

NHGRI Sequencing Budget and Annual Sequencing Capacity

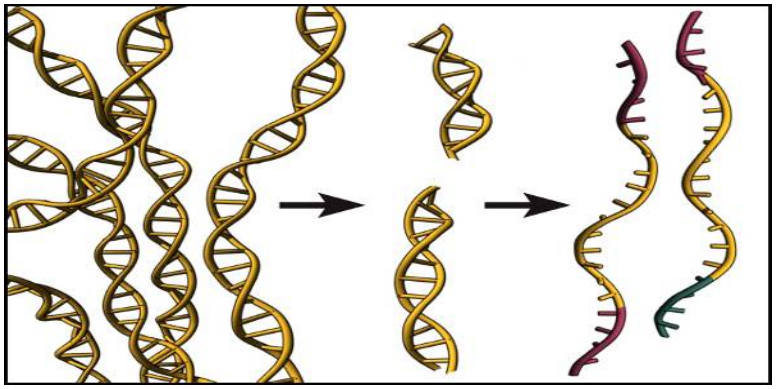


11-June-05

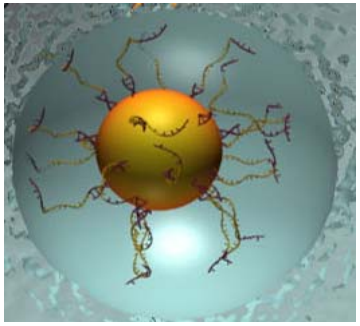
Medical Sequencing Opportunities

- Mendelian conditions
  - Rare diseases with wide linkage region
  - X-linked conditions
- Common diseases and quantitative traits
  - Convincing linkage region but no association
  - Attractive candidate genes
- Somatic mutations and cancer

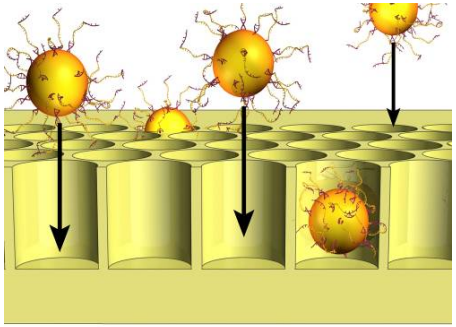
454 Sequencing Instrument



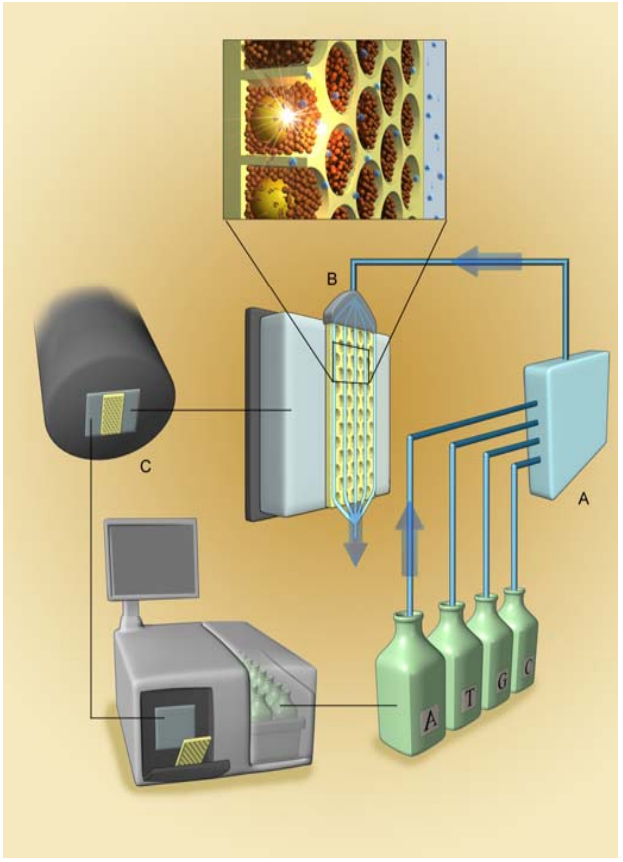
1) Prepare Adapter Ligated ssDNA Library



2) Clonal Amplification on 28 µ beads

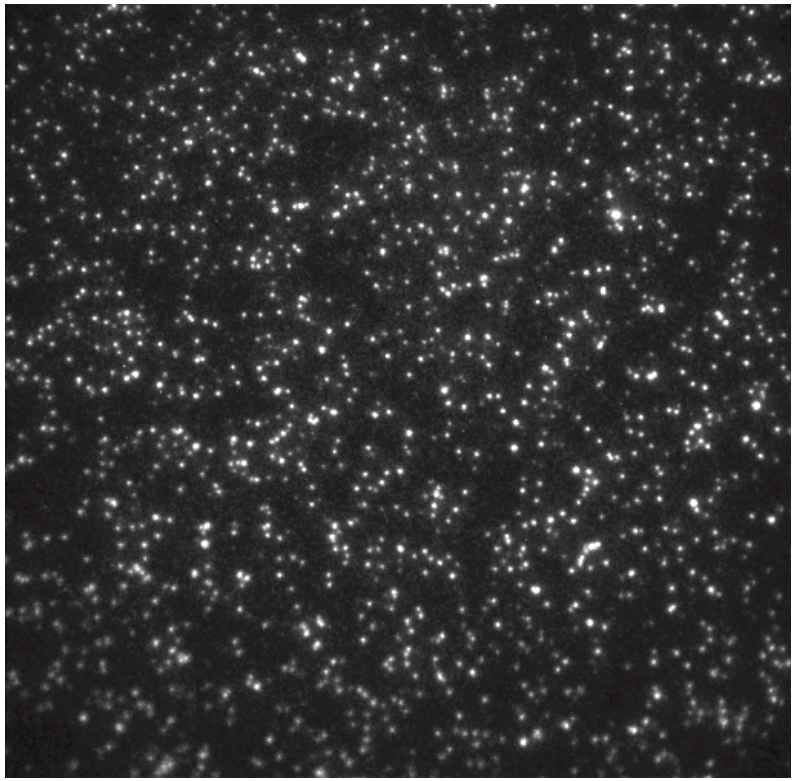


3) Load beads and enzymes in PicoTiter Plate™



4) Perform Sequencing by synthesis on the 454 Instrument

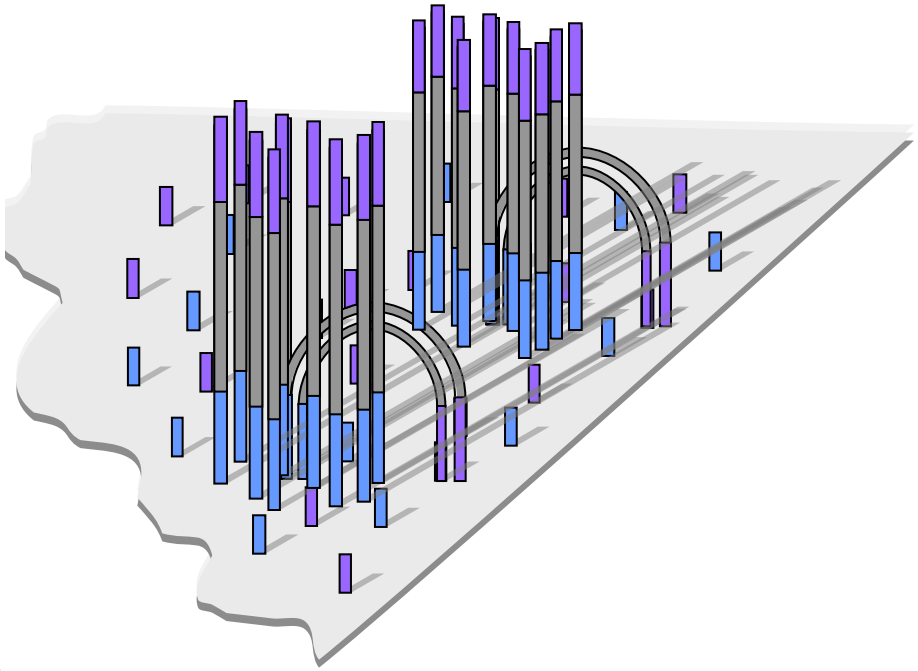
Solexa: Clonal Single Molecule Arrays™



100um

Random array of clusters

Attach single molecules to surface
Amplify to form clusters



1000 molecules per ~ 1 um cluster
1000 clusters per 100 um square
40 million clusters per experiment

As for the future, your task is not to foresee, but to enable it.

Antoine de Saint-Exupery