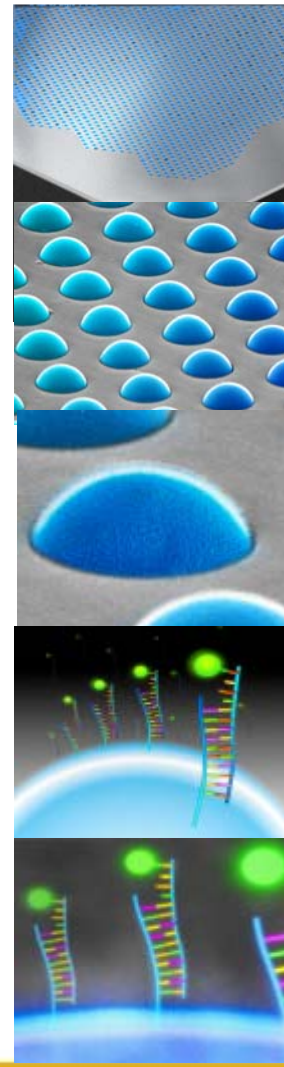
illumina®

New Developments Supporting Whole Genome Association Studies: Capturing Human Variation on a Single Microarray

David Barker, PhD
VP and Chief Science Officer



Corporate Summary

- Founded: May 1998
- IPO: July 2000
- Headquarters: San Diego, CA
- Illumina East Wallingford, CT
- Employees: >400 worldwide
- Commercialization: US/Canada, Europe, Japan, Singapore, China
  Distributors in Korea, Taiwan, Thailand, Australia)

illumina®

Topics

- Illumina microarray technology
- Lessons from The International HapMap Project
- Advances in Whole Genome Genotyping
  - Capturing all common human variation on a single microarray

illumina

Illumina Microarray Technology

illumina®

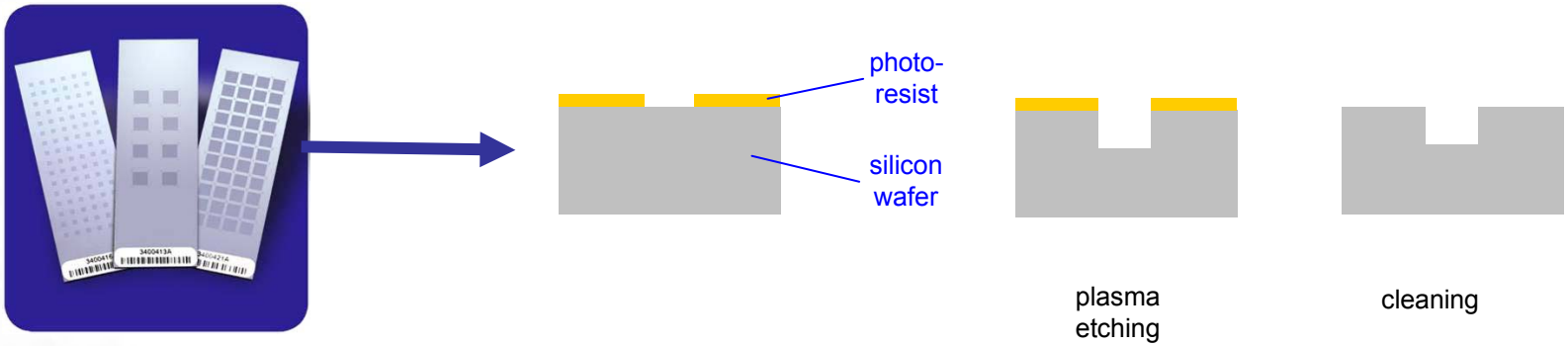
Array Formats

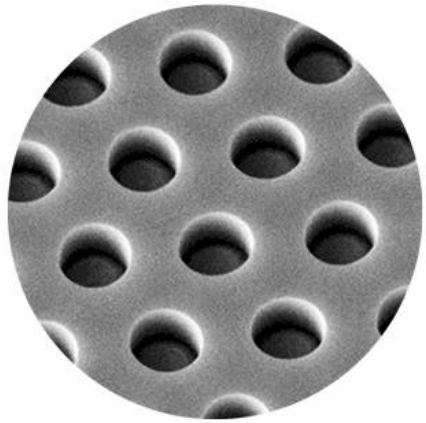
Sentrix® Array Matrix

Sentrix BeadChips

illumina®

Microfabrication of Wells

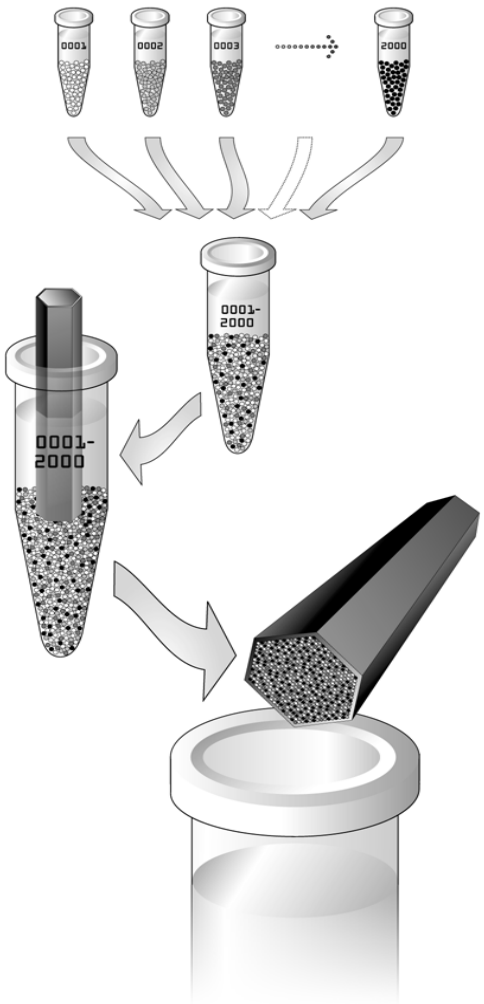




Microwells

illumina

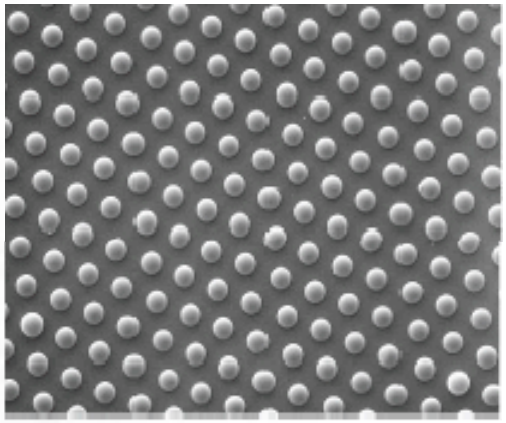
Bead Preparation and Array Production

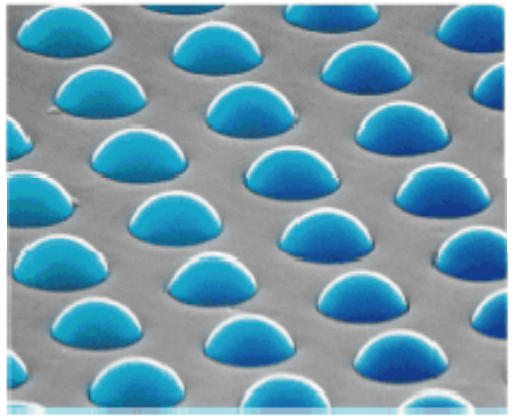


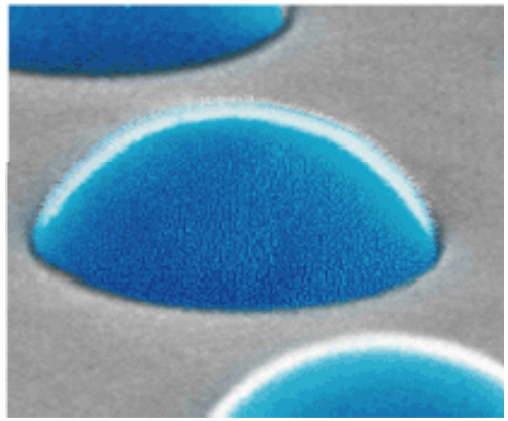
- Beads individually manufactured and QC'd with full length Oligator oligos
- Bead pools produced containing 96 to 60,000 bead types
- Bead pools applied to fiber bundle or BeadChip; beads self-assemble into wells to form functional array. *Average of 30 beads of each type.*
- All elements of the array quality checked prior to supply (Gunderson et al. *Decoding randomly ordered DNA arrays.* Genome Research, May 2004)

illumina

Beads in Wells

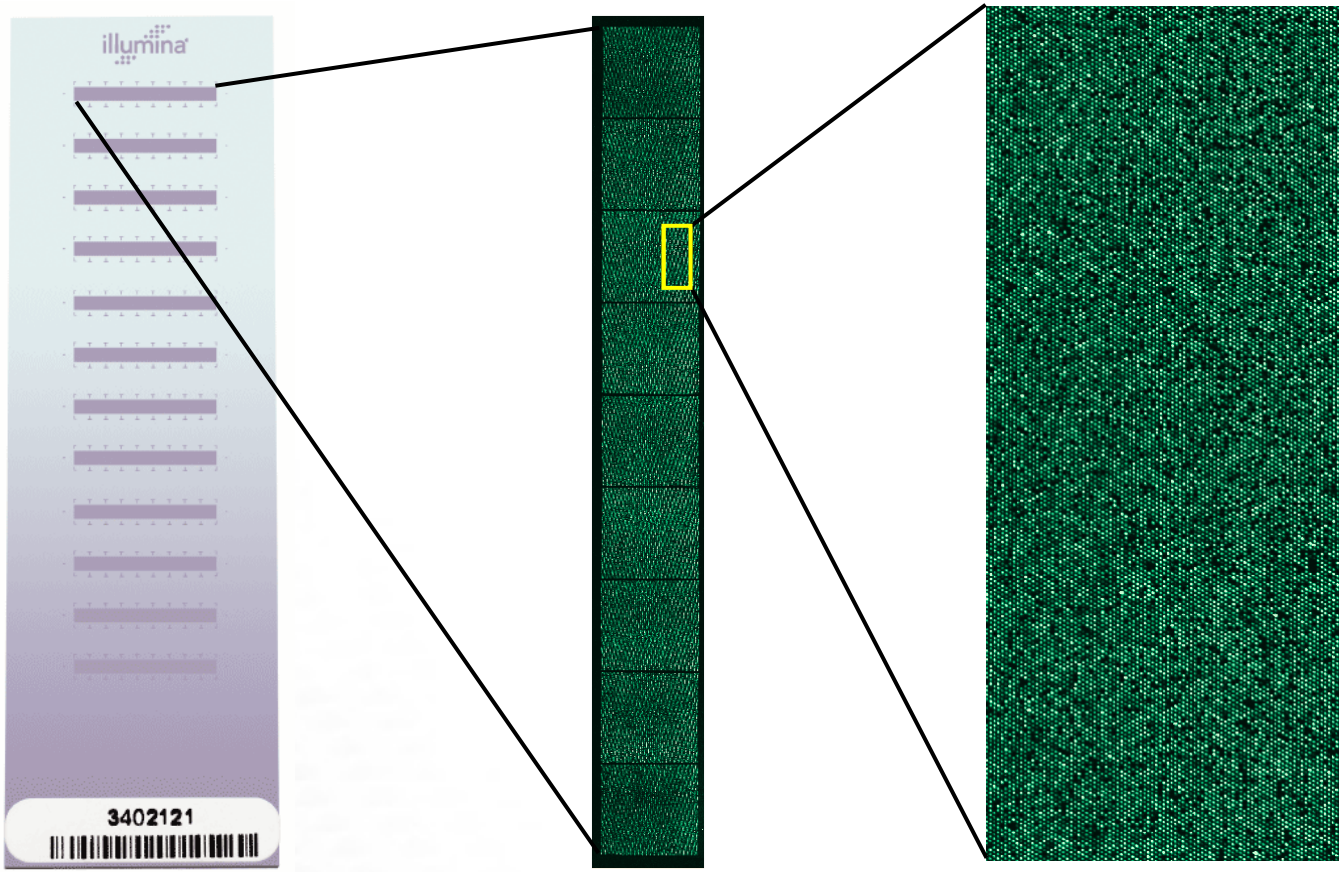






illumina®

Achieving High Density Genotyping



10-20 Sections
(30 K bead types = 30K SNPs)

>1.2M features
per section

Average 20 fold
redundancy

illumina

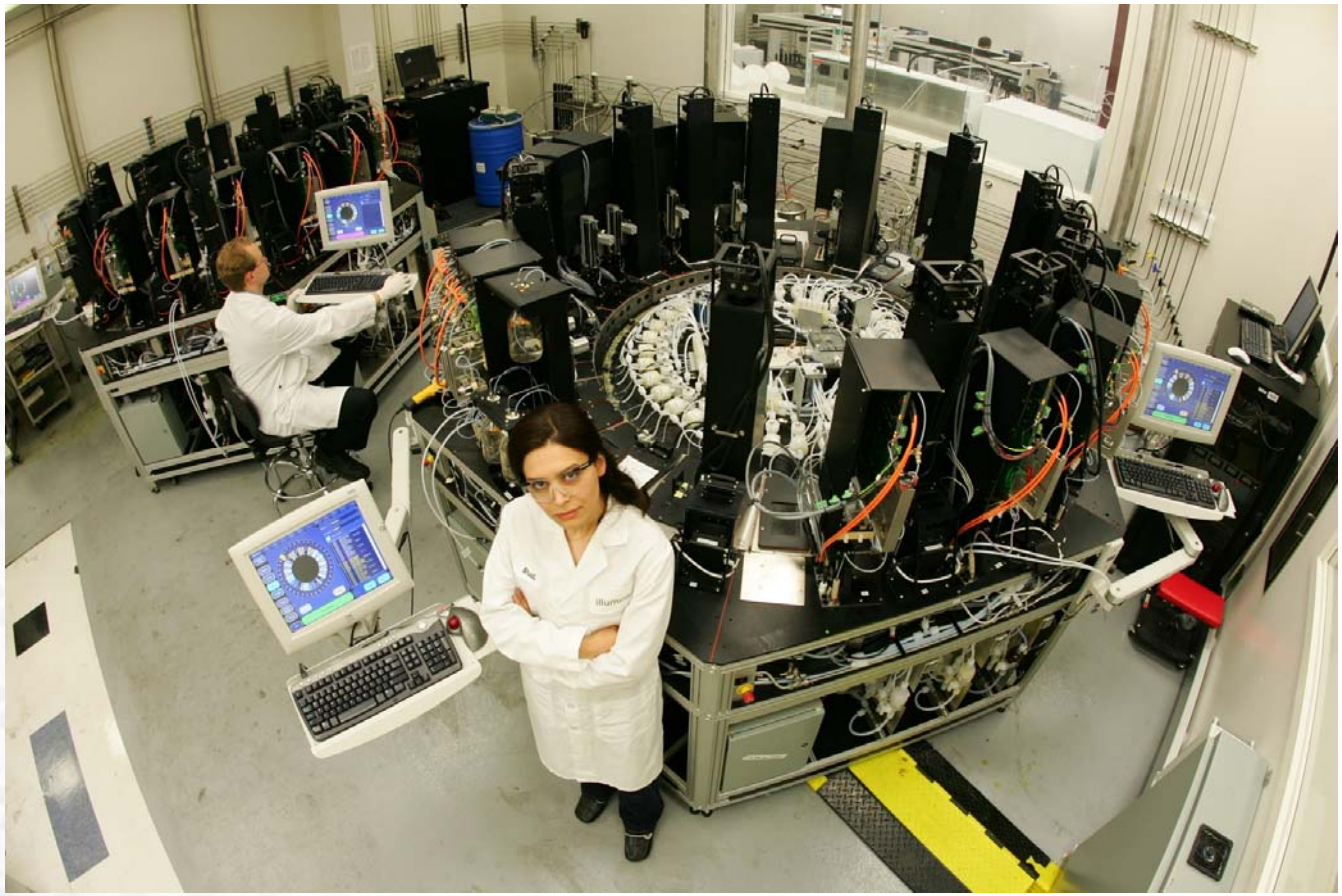
BeadStation
Powerful & Economical BeadArray Platform

- Confocal scan with two lasers
- Collects two colors simultaneously
- Resolution = 0.8 microns

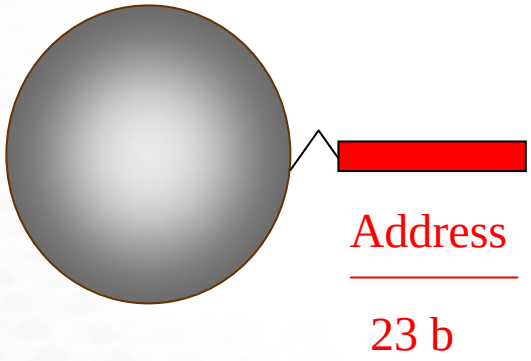
illumina

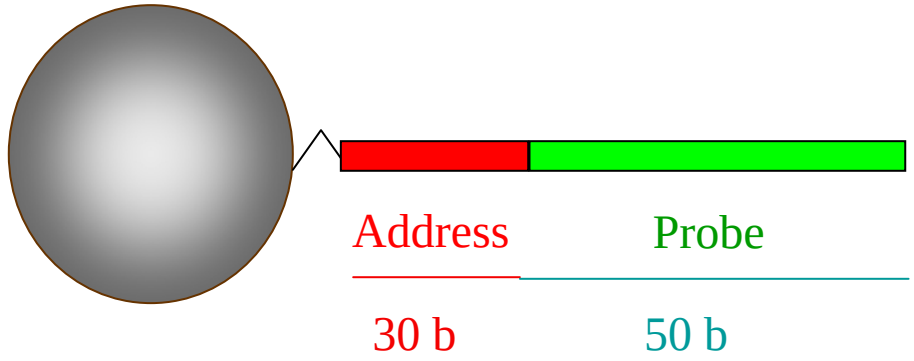
Oligator Synthesis Factory

- Microarrays and assays need lots of oligonucleotides
- New 4th generation synthesizers can supply worldwide oligo needs
  - 13,824 oligos per run; 2 million bases/day



Illumina's Bead Designs





- GoldenGate® Genotyping, DASL™ gene expression, allele-specific expression and methylation analysis
  - Universal array
  - 23-base address code
- Infinium™ genotyping and gene expression with IVT labeling
  - 30-base address code
  - 50-base sequence-specific probe linked to address

illumina®

Lessons from the HapMap Project

illumina®

Common Disease, Common Variant Hypothesis

- The genetic component of common diseases results from the simultaneous occurrence of several contributing gene variants
- These gene variants are relatively common in the human population
- Working definition: “Common” means ≥ 5% of people have the variant (M.A.F ≥ 0.05)
- The challenge: discover these sets of common variants that result in greater risk for disease
- The hope: knowing the genes involved will provide scientific understanding and targets for therapy

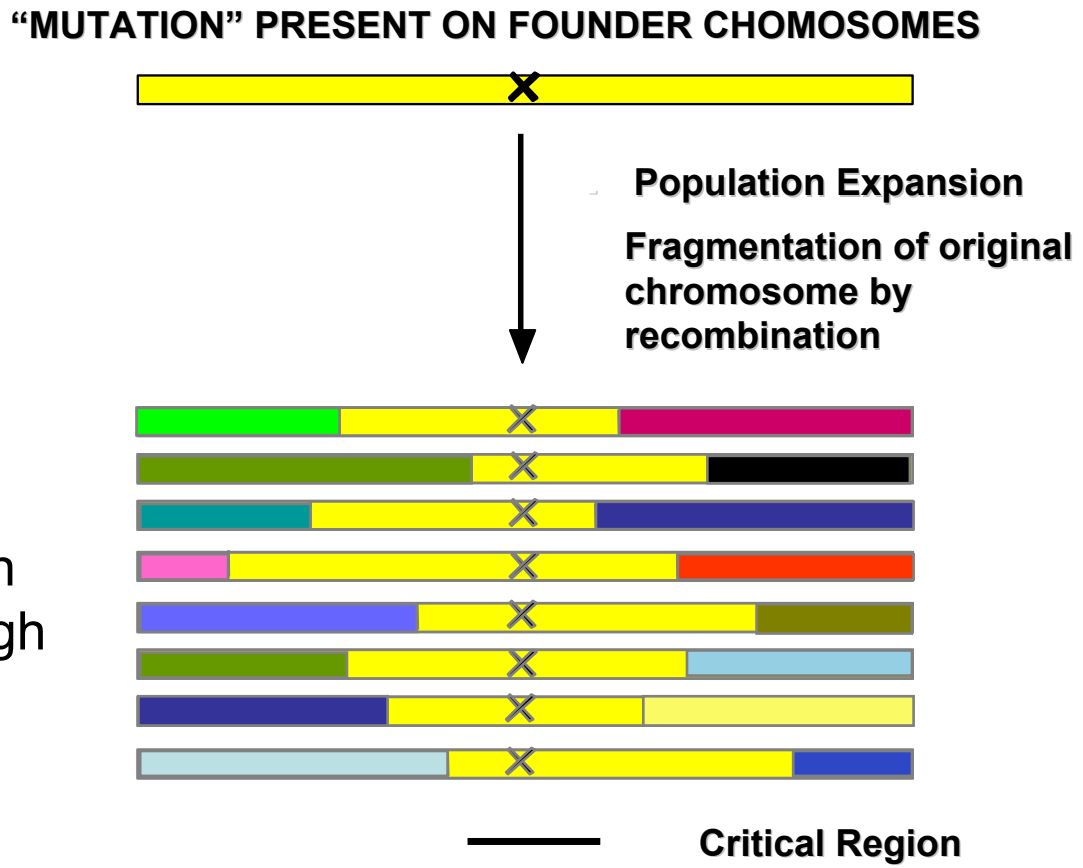
SNPs as Markers of Genomic Variation

- “The” human genome sequence is known
- How to we study genome variation?
- Single Nucleotide Polymorphisms (SNPs)
  - Very abundant—over 10 million in the human genome
- Can SNPs be used to study variation genome-wide?
  - Need efficient and inexpensive method for assay
    - Cost in 2000: ~$1 per SNP

Understanding the History of Variations

By typing SNPs across the genome, critical regions shared by disease carriers can be defined.

SNPs near the disease mutation will be associated with it with high statistical significance.



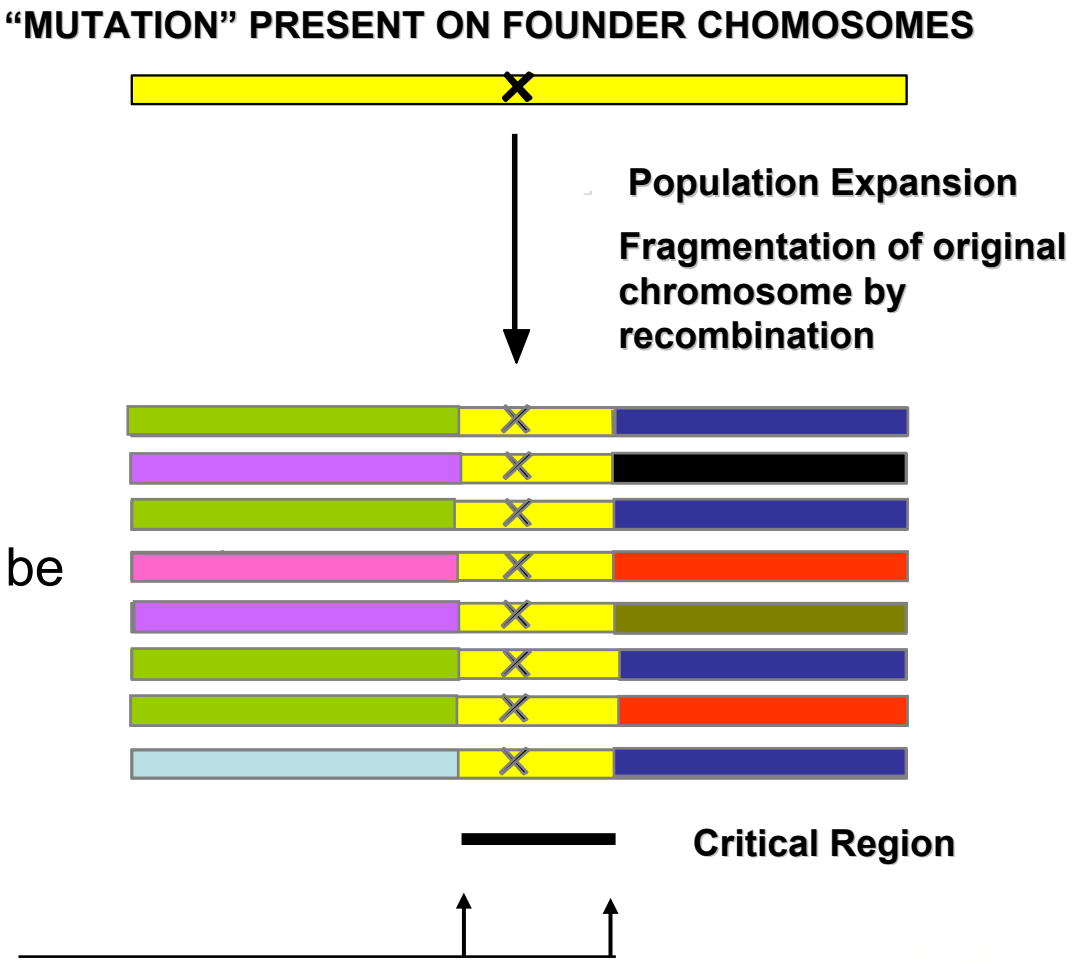
illumina

Is Variation "Block-like" Rather than Random?

Early association evidence suggested that the genome exists in blocks (haplotypes) of inherited material, of limited diversity.

A few SNPs in each block may be able to represent the rest.

"Hot spots" of recombination?



illumina

www.hapmap.org

Objectives of the HapMap Project

- Describe the common patterns of sequence variation in the human genome
- Include multiple populations with ancestry from parts of Africa, Asia and Europe
- Make this information freely available in the public domain
- Provide tools to aid discovery of genetic variants that affect common disease in association studies

HapMap Participants and Contribution: Phase 1

Country	Research Group and Leaders	Chromosome Assignment	% of Genome
Japan	Yusuke Nakamura, RIKEN/University of Tokyo	5, 11, 14, 15, 16, 17, 19	25.1%
United Kingdom	David Bentley, Wellcome Trust Sanger Institute	1, 6, 10, 13, 20	24.0%
Canada	Thomas Hudson, McGill University	2, 4p	10.0%
China	The China HapMap Consortium	3, 8p, 21	10.0%
United States	Illumina Whitehead Baylor UCSF & Washington Uni	8q, 9, 18q, 22, X 4q, 7q, 18p, Y 12 7p	15.5% 9.1% 4.4% 1.9%

- 1,000,000 SNP assays developed
- 285,000,000 genotype calls made
- ~70% done with Illumina GoldenGate assay

illumina

Vol 437|27 October 2005|doi:10.1038/nature04226

nature

ARTICLES

A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.

Lessons from the HapMap Project

- Haplotype block structure is significant
  - Size of blocks in CEU and CHB/JPT samples is similar
  - Average size of blocks is much smaller in African population
- Use of well chosen “tag” SNPs will facilitate whole genome association analysis
  - Need only ~1/3 as many tag SNPs as random SNPs
- From 100,000 to 600,000 SNPs should be enough to represent all common human variation
  - ~100,000 in isolated populations
  - ≥600,000 in very old, diverse populations

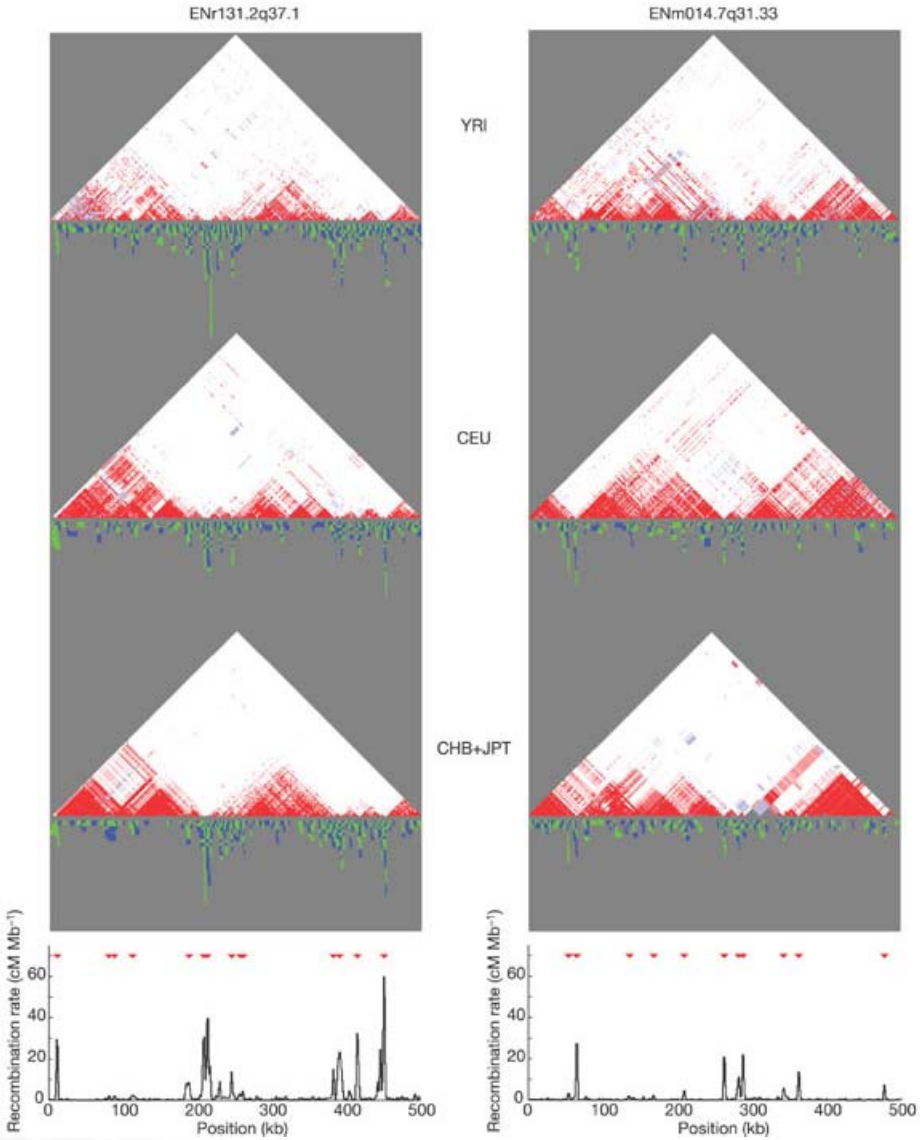
Haplotype Block Structure in the Human Genome

Parameter	YRI	CEU	CHB+JPT
SNPs per block	30.3	70.1	54.4
Ave. block length (kb)	7.3	16.3	13.2
Fraction of genome in blocks (%)	67	87	81
Ave. haplotypes per block (MAF>0.05)	5.57	4.66	4.01

Nature **437**: 1299 (2005) Table 5

illumina

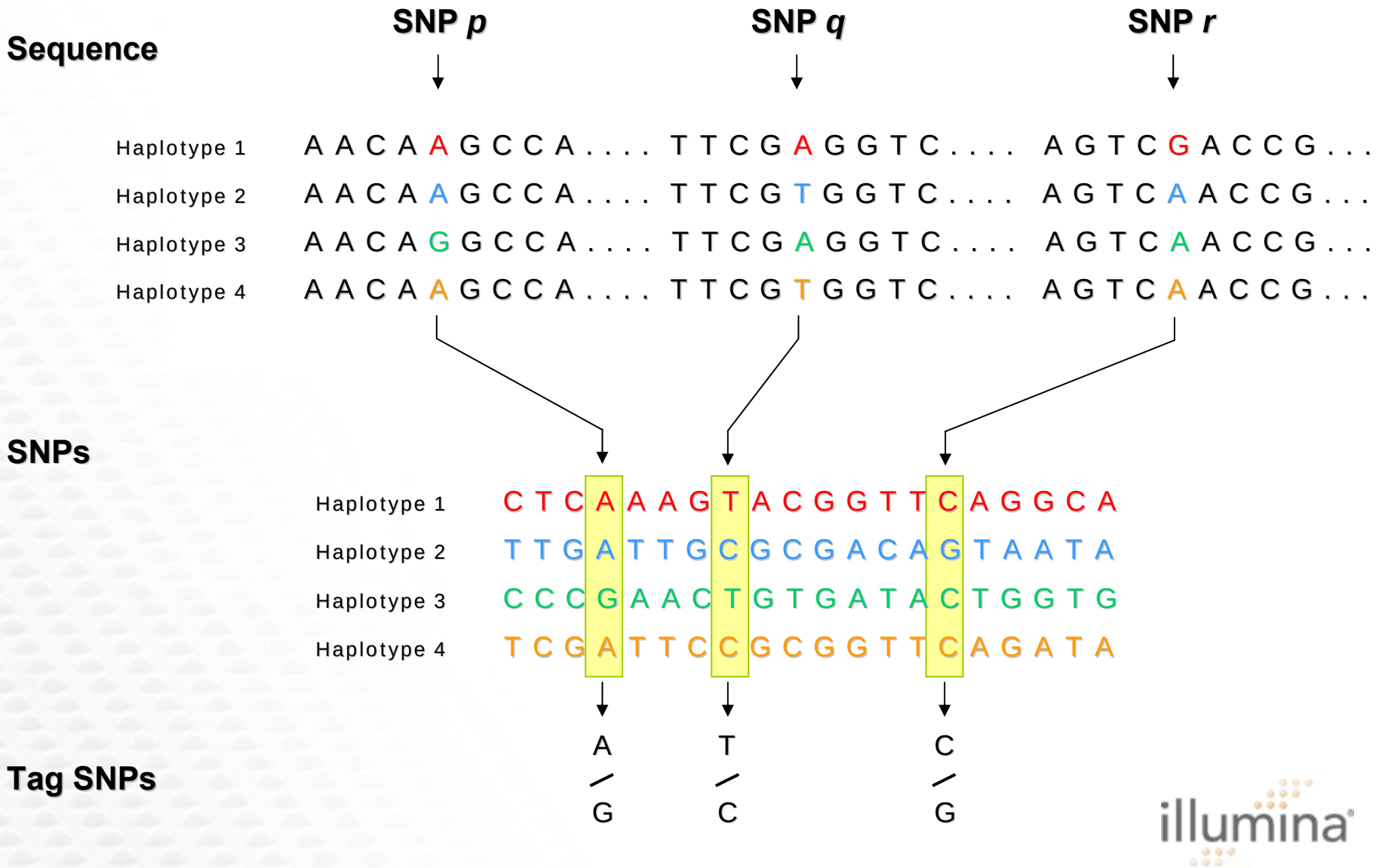
Haplotype Blocks



Nature **437**:1299 (2005) Figure 8

illumina

Picking Haplotype Tag SNPs

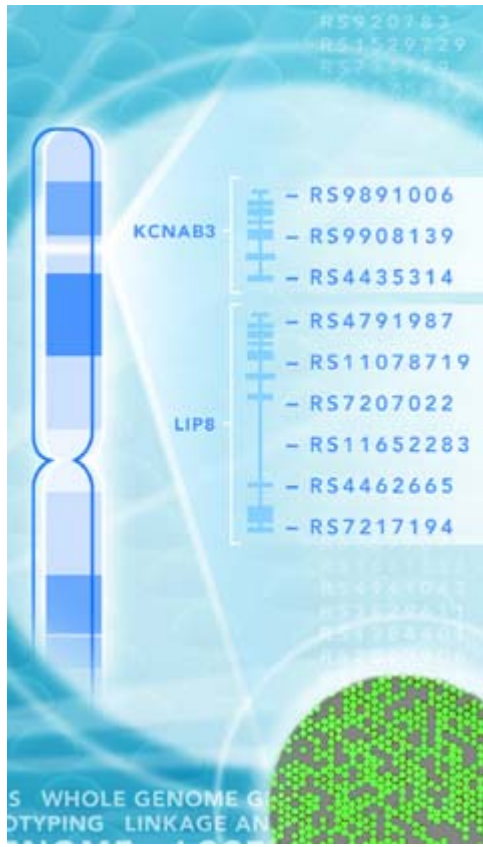


illumina

Whole Genome Association Studies

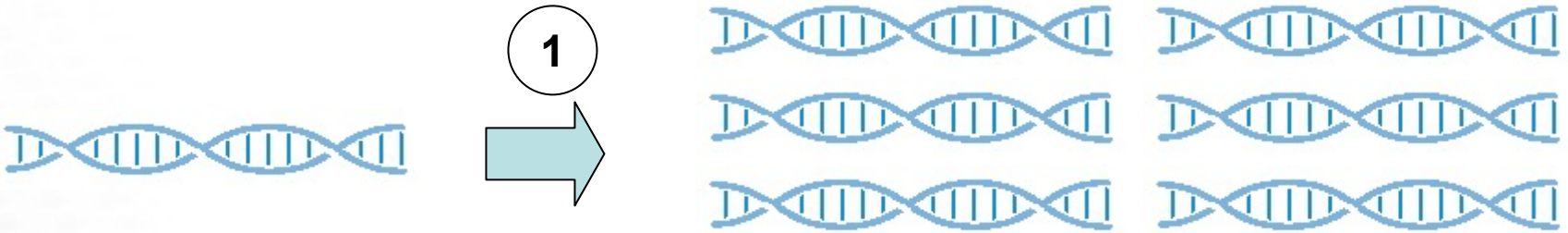
What are the key needs?

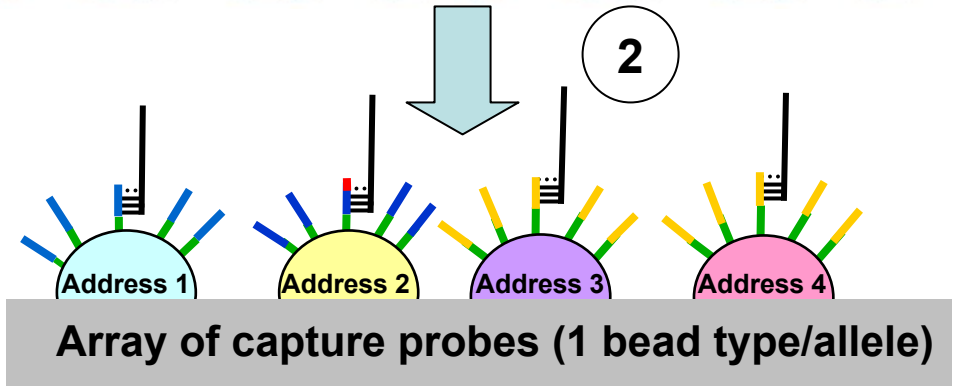
- Genotype 100,000's of loci accurately
  - High locus selectivity
  - High specificity for allelic discrimination
- Ability to assay tag SNPs; access to vast majority of genome
- A robust means of processing many samples easily and efficiently
- A technician-friendly automatable process that reduces possibility of sample tracking error

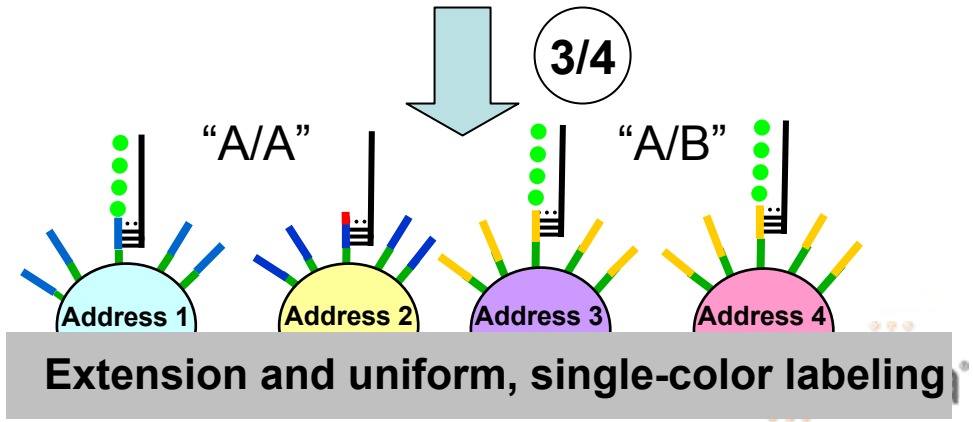


illumina

The Infinium Assay

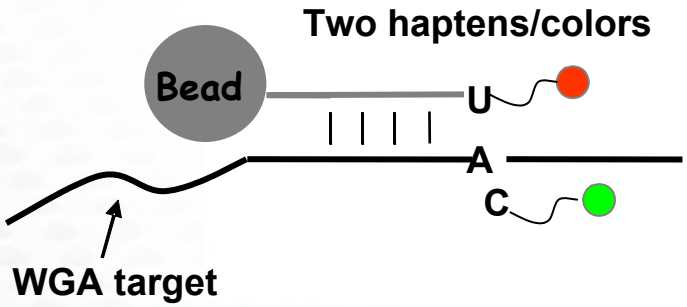




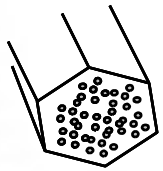


1. Amplify genome uniformly
2. Fragment, Denature and Hybridize to immobilized 50-mers
3. Discriminate SNPs (enzyme-mediated)
4. Amplify signal and read out SNP calls

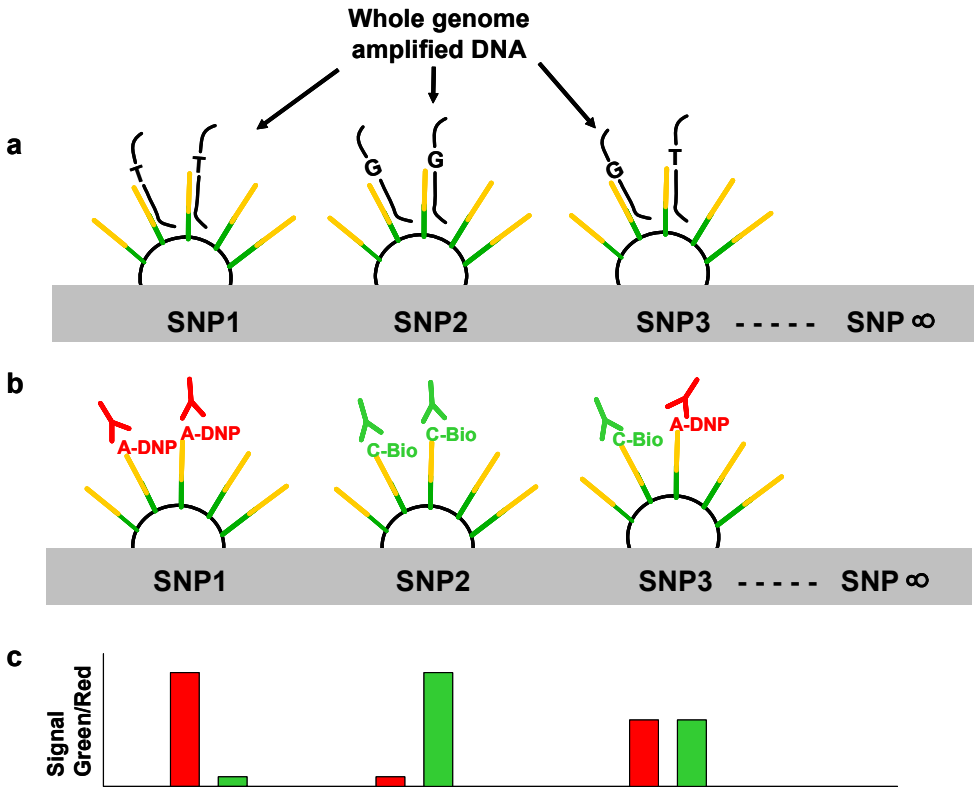
Infinium II Assay
Single Base Extension



ddA,ddC,
ddG,ddU

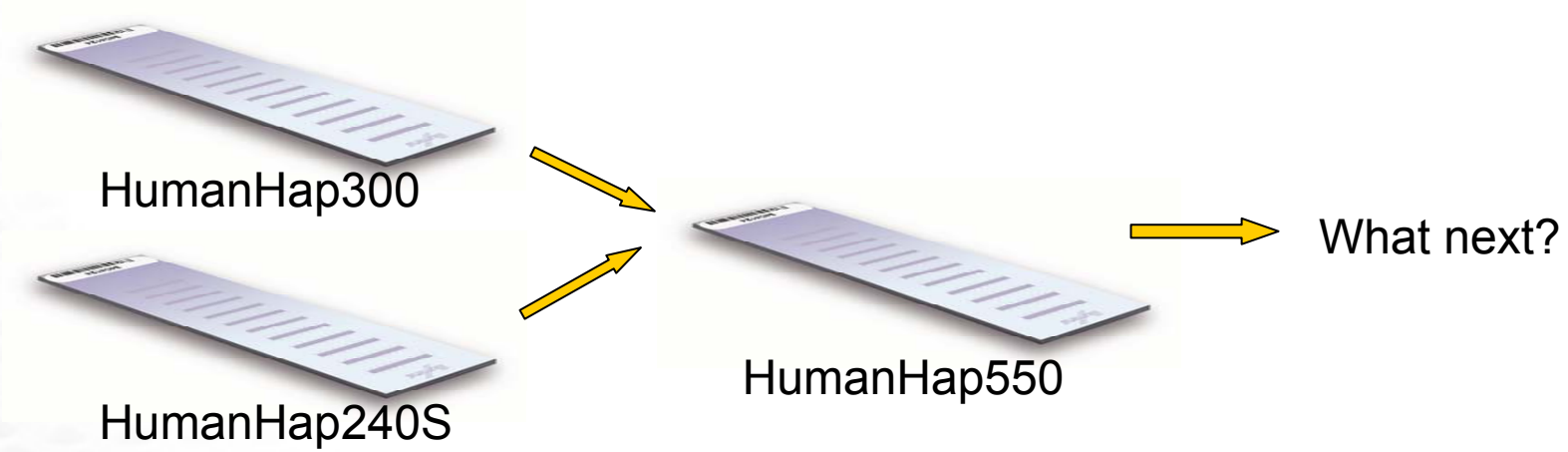


{[A/G], [C/T], [A/C], [G/T]}



illumina

Infinium: Efficient Assay of Human Variation



More than 550,000 SNPs assayed on one chip

illumina

HumanHap550 Tag SNP Strategy

- Analyze full HapMap data set (Phase I + II) using tools to select tag SNPs
- Association content strategy
  - $r^2 \geq 0.80$ for large bins (≥ 3 SNPs) in CHB+JPT population
  - $r^2 \geq 0.80$ for bins containing SNPs within 10kb of genes or in evolutionarily conserved regions (ECRs) in CEU
  - $r^2 \geq 0.70$ for large bins (≥ 5 SNPs) in YRI population
- What does r^2 mean?
  - Twice as many samples needed for $r^2 = 0.5$ than if $r^2 = 1.0$

illumina

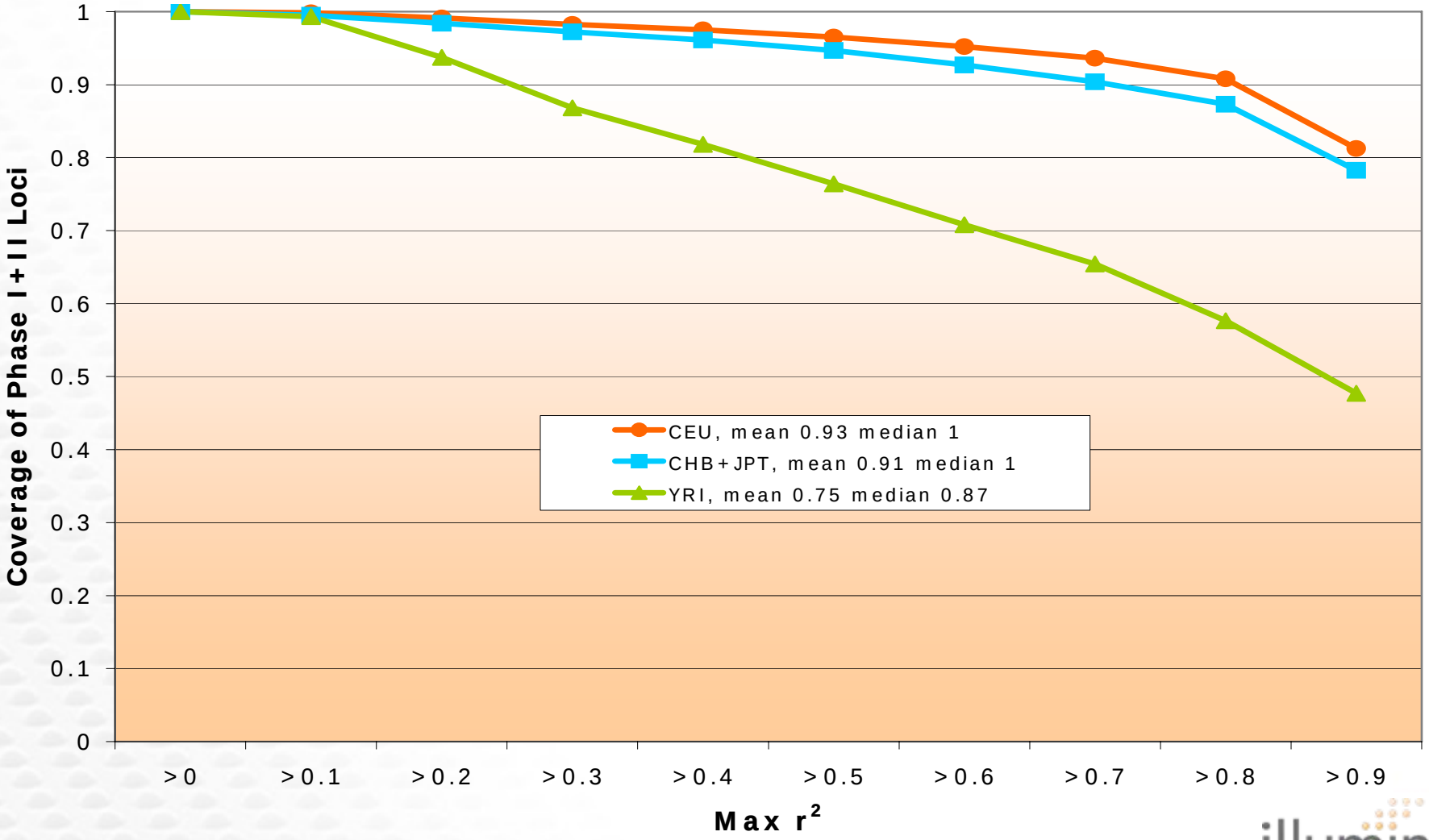
HumanHap550 SNP Content

- **Tag SNP content:**
  - > 550,000 tagSNPs chosen from the HapMap Project.
    - >317k tag SNPs chosen from the Phase I HapMap project
    - 240k additional tag SNPs chosen from Phase II HapMap Project
      - Substantially increased coverage in all populations, especially Han Chinese/Japanese and Yoruba

- **Additional content:**
  - 7,300 nsSNPs
  - 1,200 MHC SNPs
  - 4,300 SNPs chosen from reported copy number regions of the genome
  - 160 mitochondrial SNPs
  - 800 ancestry informative markers (AIMs)

illumina

HumanHap550: Coverage by Population



illumina®

HumanHap300 Data Quality

127 samples
25 trios
15 replicates

Parameter	SNP Counts	Percent	Product Specification
Call rate	40,295,386 / 40,322,881	99.93%	>99% (average)*
Reproducibility	4,760,834 / 4,760,893	>99.99%	>99.9%
Mendelian Inconsistencies	2,610 / 7,919,883	0.033%	<0.1%
Concordance with HapMap Data	33,776,528 / 33,847,060	99.79%	

illumina

HumanHap240S Data Quality

117 samples
27 trios
5 replicates

Parameter	Counts	Percent	Product Specification
Call rate	28,278,645 / 28,302,956	99.91%	Average >99%**
Reproducibility	1,218,112 / 1,218,113	>99.99%	>99.9%
Mendelian Inconsistencies	2,063 / 6,454,342	0.032%*	<0.1%
Concordance with HapMap Data	25,708,175 / 25,789,740	99.68%	

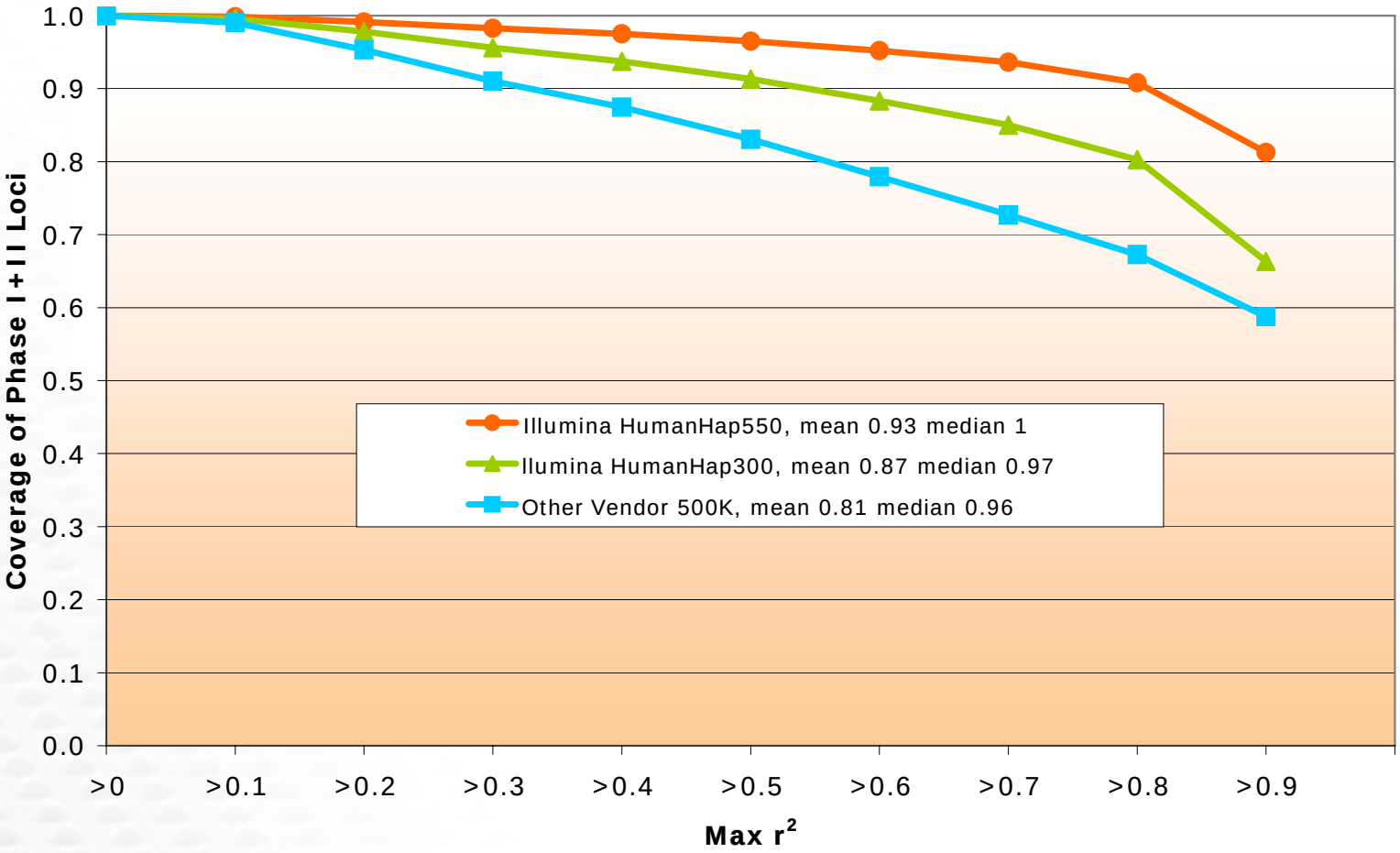
illumina

HumanHap300 Quality Control and Performance Statistics from CIDR Pilot Project Data (FUSION study)

Loci	317,503
Samples	946
Average Call Rate	99.8%
Reproducibility	99.9998%
Mendelian Trio Consistency	99.99%
Concordance with Hapmap Genotypes (18 CEPH samples)	99.85%

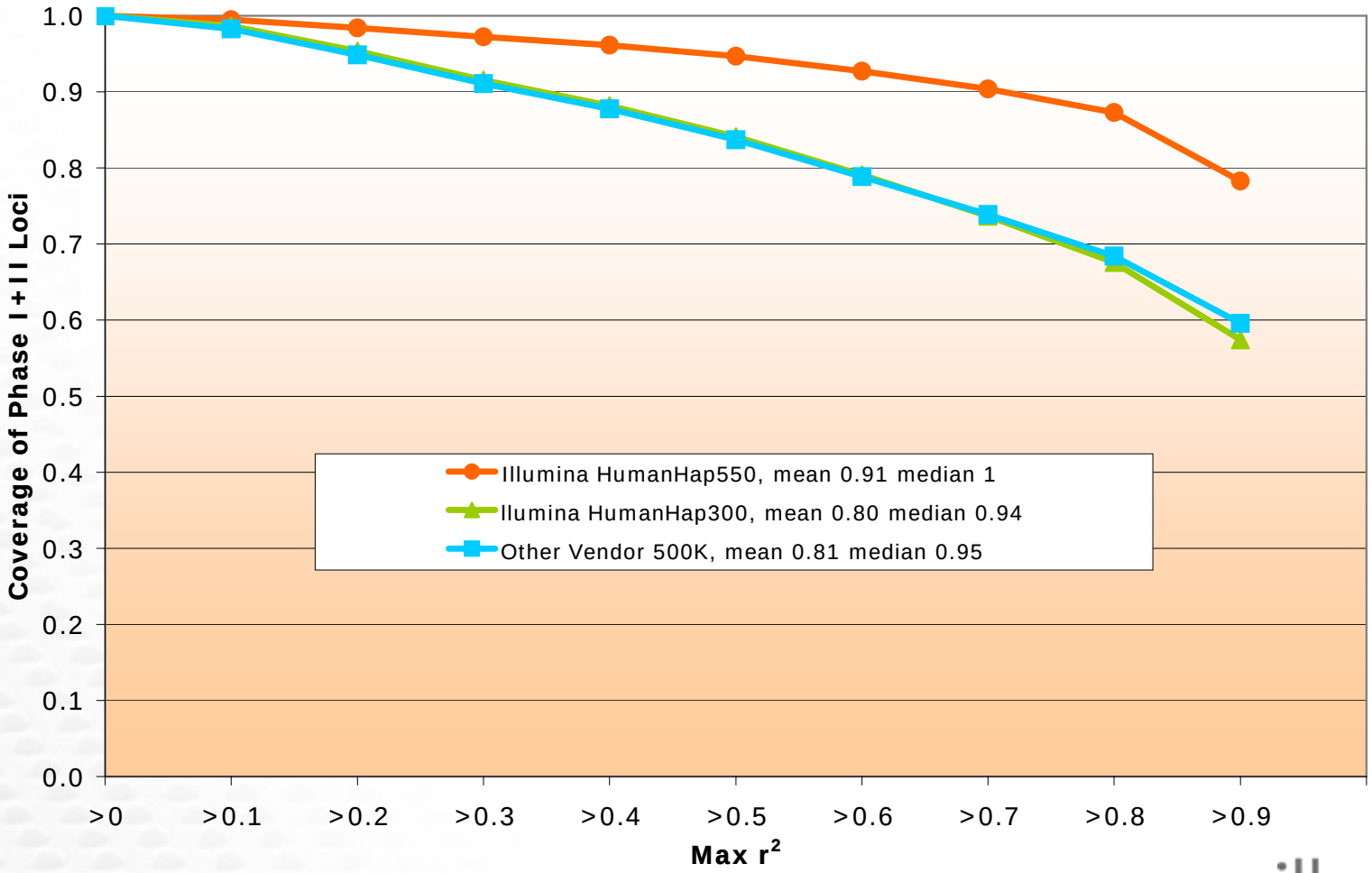
illumina

Comparison of Whole Genome Genotyping Panels
CEU Population, MAF ≥ 0.05



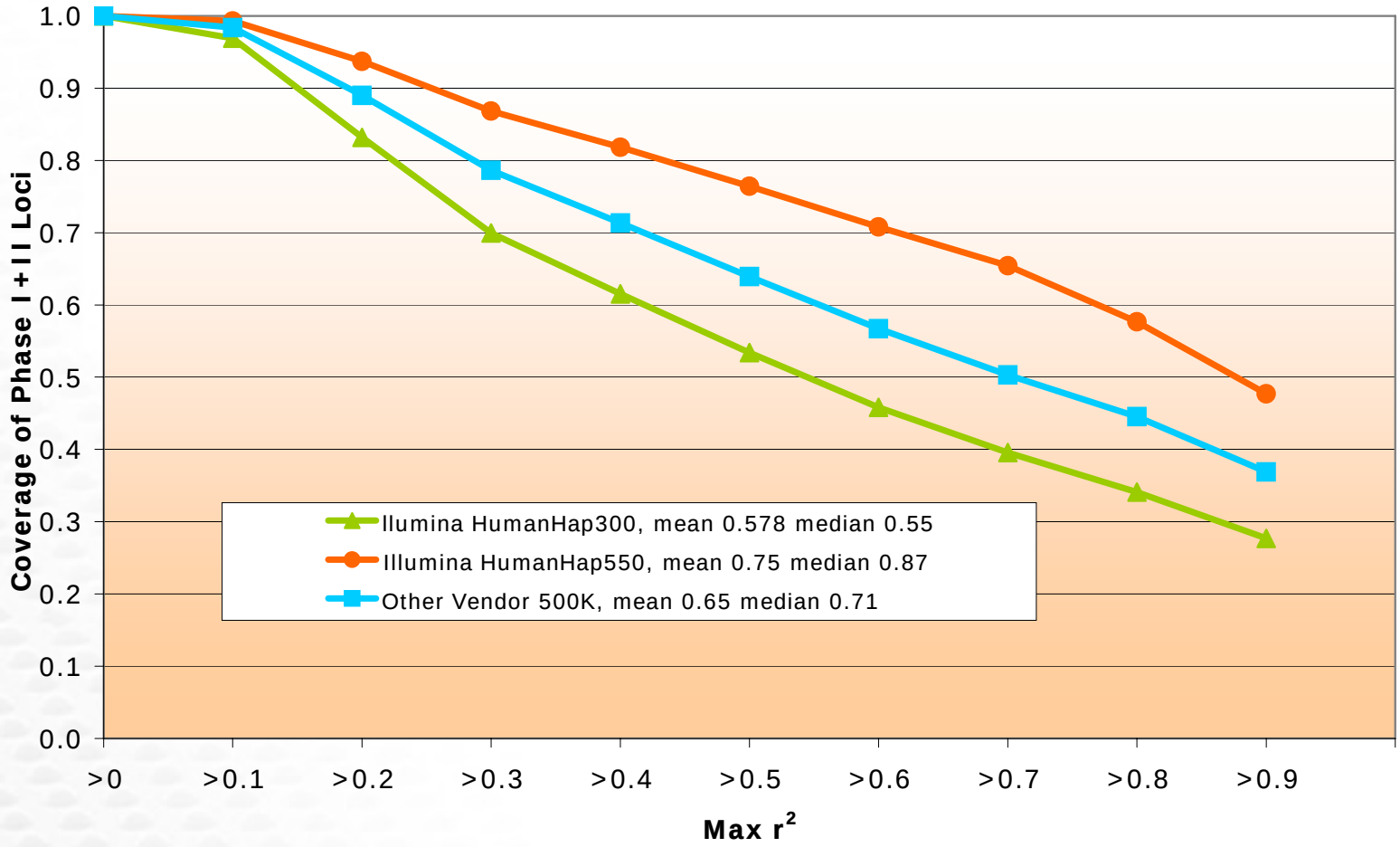
illumina®

Comparison of Whole Genome Genotyping Panels CHB+JPT Populations, MAF ≥ 0.05



illumina

Comparison of Whole Genome Genotyping Panels
YRI Population, MAF ≥ 0.05



illumina

Ancestry Informative Markers

- AIMs are important to assess potential population stratification between case and control groups in an association study

- DNA Test Panel (280 SNPs):
  - SNPs chosen to have highly signficant allele frequency differences in at least one possible pairwise combination using HapMap populations
    - Yoruba - Han Chinese/Japanese
    - Yoruba – CEU
    - Han Chinese/Japanese – CEU

- AIMs markers selected by Seldin et al. (500 SNPs):
  - Northern and Southern European subgroups

Mitochondrial SNPs

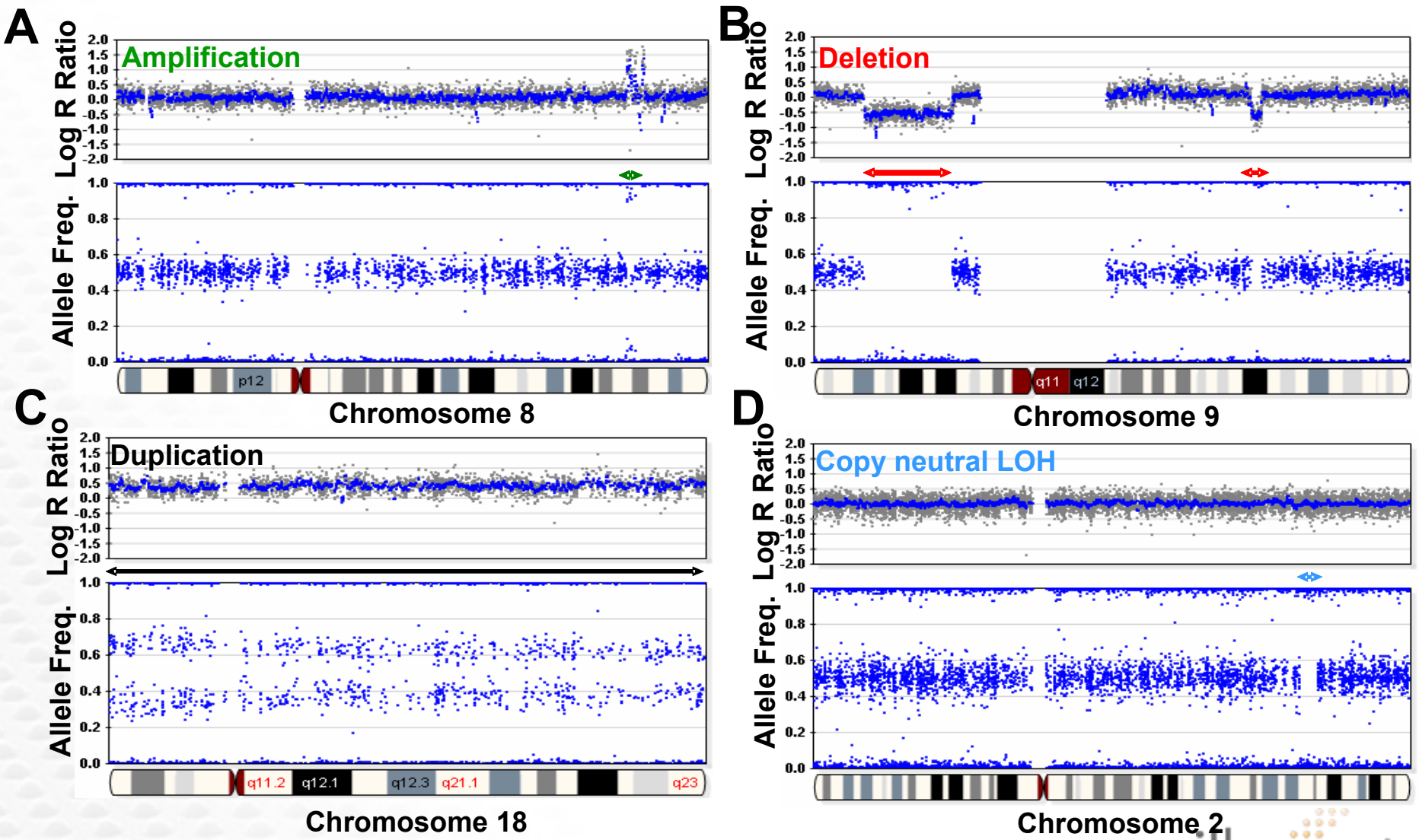
- Mitochondrial genome contains genes involved in metabolism and genes implicated in disease (i.e., Parkinson's Disease)
- Selection criteria: MAF ≥ 0.01 in European population, obtained from Broad Institute (http://www.broad.mit.edu/mpg/tagger/mito.html)
- HumanHap550 is the only whole-genome genotyping product that contains mitochondrial SNPs
- 180 SNPs chosen; subsets polymorphic in each population
  - 46 SNPs in 33 unrelated CEU individuals
  - 48 SNPs in 29 unrelated CHB/JPT individuals
  - 60 SNPs in 26 unrelated YRI individuals

illumina

SNPs in Known Copy Number Polymorphism Regions

- Regions identified from 3 recent publications
  - Conrad, et al. (Nature Genetics 38:75-81, 2006)
  - Hinds, et al. (Nature Genetics 38:82-85, 2006)
  - McCarroll, et al. (Nature Genetics 38:86-92, 2006)
- 4,317 SNPs chosen and included in HumanHap240S
- CN regions detected by loss of heterozygosity (LOH)
- 495 regions from the 3 publications confirmed
- Evidence that SNPs are in LOH (CN) regions:
  - 20X higher Mendelian Inconsistency rate compared to rest of HumanHap240S SNPs
  - Observed lower CN scores in BeadStudio software in these defined regions

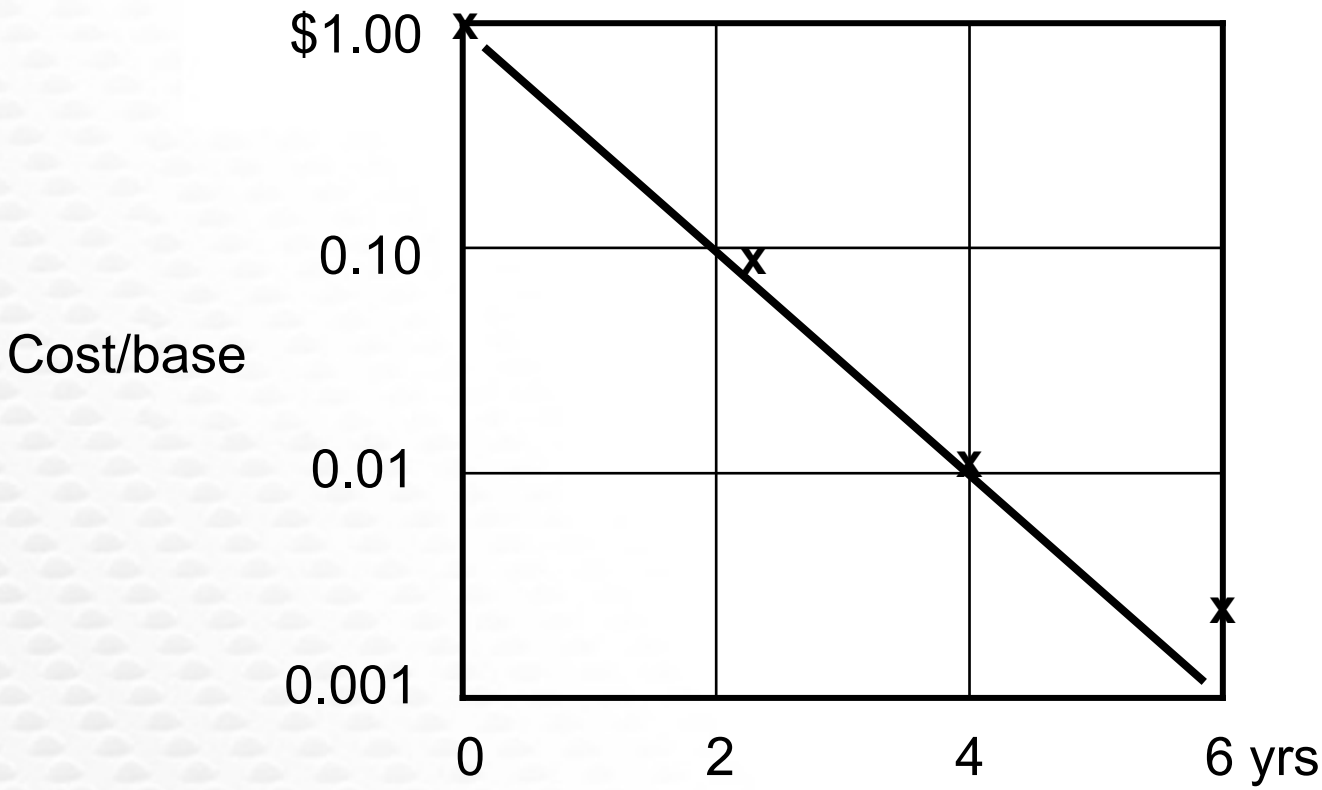
BeadStudio Analysis of Copy Number Variations



illumina

Cost of Genotyping

- In 2000: ~$ 1 per SNP
- HumanHap550
  - ~$0.0025 per SNP
  - 400-fold reduction in price in 6 years



illumina

Conclusions

- Human variation can be studied efficiently on BeadArrays
  - High-density platform
  - >550,000 SNPs
  - High accuracy
  - Low cost
- Gold rush to find associations
  - Understanding complex diseases
- Should lead to the era of personalized medicine

illumina

Acknowledgements: A Team Effort

Assay Development

Chan Tsan
Joe Musmacker
Patrick Merrit
Dave Bullis
Hongji Ren
Ken Kuhn
Richard Shen
Weihua Chang
Frank Steemers
Kevin Gunderson

Bioinformatics

Pauline Ng
Lixin Zhou
Jim Bierle
Bryan King
Semyon Kruglyak
Bahram Kermani
Francisco Garcia

Manufacturing

Eric Johnson
Mark Staebell
Celeste McBride
Dave Douglas

Bead Loading

Lea Perez
Paul Kitabjian
Melissa Wiley
Rob Yang
Jerry Zhou
Ryan Smith
Steve Barnard
Michael Graige

Chemistry R&D

Igor Koslov

Gali Steinberg
Chanfeng Zhou
Michael Lebl

Oligator Team

Aaron Jones
Steve Burgett
Mark Nibbe
Igor Koslov
Michael Lebl
Tom Rosso
Ken Sikes
David Heiner

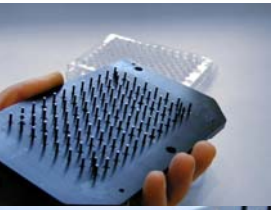
Marketing

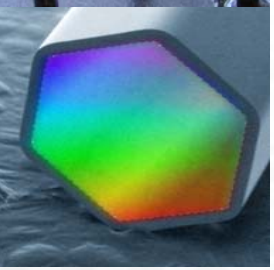
Sarah Murray
Vivian Zhang
Todd Dickinson

Sr. Staff

John Stuelpnagel
David Barker
Bob Kain
Scott Kahn

illumina



illumina®

Products and Services for Genetic Analysis