

Utilizing Human Variation in the Human Genome in Molecular Epidemiology of Tobacco, Diet and Cancer

Stephen Chanock M.D.,

Senior Investigator, Pediatric Oncology Branch,
NCI

Director, Core Genotyping Facility, NCI



<http://cgf.nci.nih.gov>

<http://snp500cancer.nci.nih.gov>

ALL RIGHTS RESERVED
<http://www.cartoonbank.com>



*“What ever will we think about now that
the genome project is almost complete?”*

SNPs, Haplotypes and Genomics

- **Clinical Outcomes**
 - **Susceptibility**
 - Gastric Cancer *IL8*
 - Bladder Cancer (tobacco) *NAT2*
 - **Disease Outcome**
 - Thrombosis *FV*
 - CGD *MPO, FCGR2A*
 - Cystic Fibrosis *MBL2*
- **Pharmacogenomics**
 - Life-threatening toxicity *TPMT in Leukemia Rx*
- **Population genetics**
 - Evolution and selective pressure *HBB*
 - Migration *CCR5*

How Do We Choose Genetic Variants for Study?

- Hypothesis-driven choice of genes
 - **Gene-environment**
 - *NAT2* and Smoking and Bladder Cancer
 - **Disease model**
 - Primary pathogenesis- *CCR5* and HIV Infection
 - Genetic modifiers- *MBL2* in Cystic Fibrosis
- Pathway of biologically related genes
 - **Known orthologs**- sequence homology between species
 - **Interacting proteins**- compliment proteins
- Scan Across the Entire Genome (no specific hypothesis)
 - Region mapped by linkage study

Folate Metabolism Pathway and ALL

Reduction in adult risk

SHMT1 (OR=0.48)

Cytosolic serine hydroxymethyltransferase

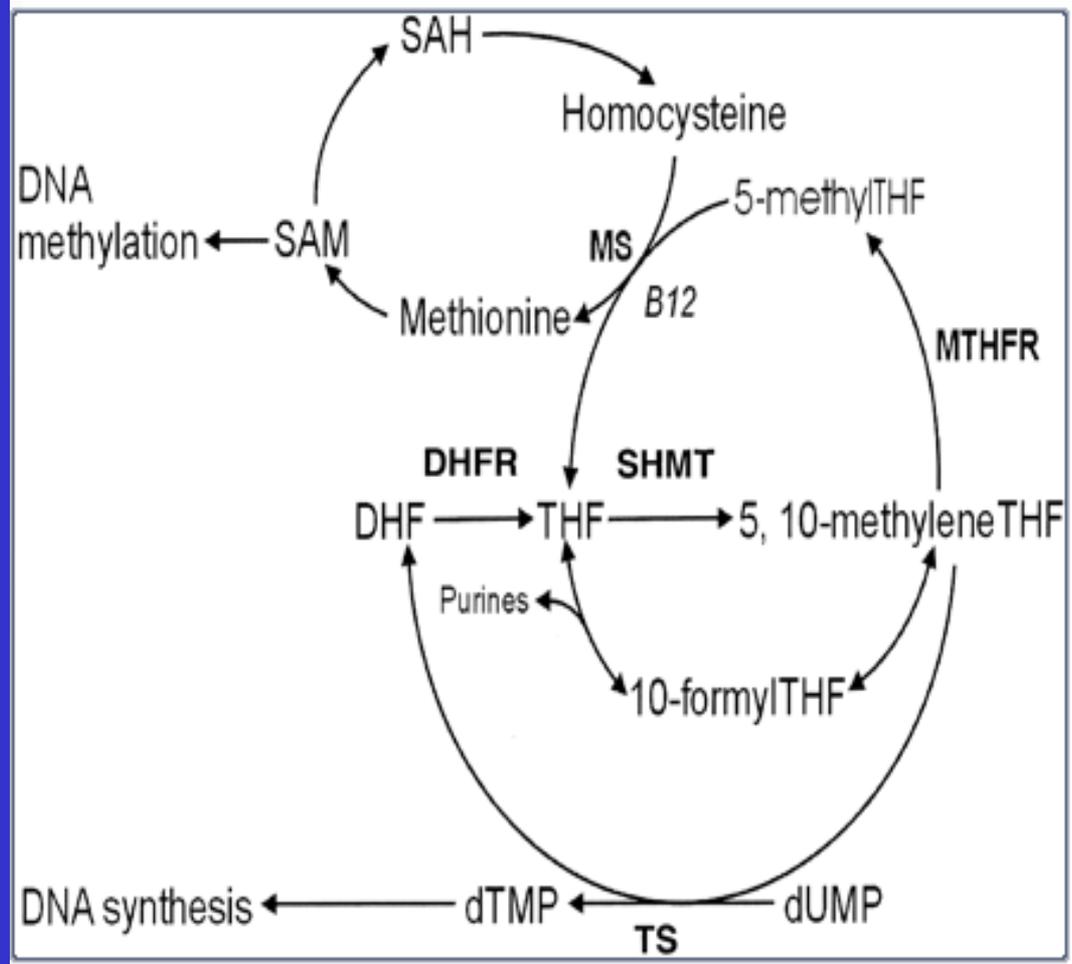
TS (OR=0.36)

Thymidylate synthase

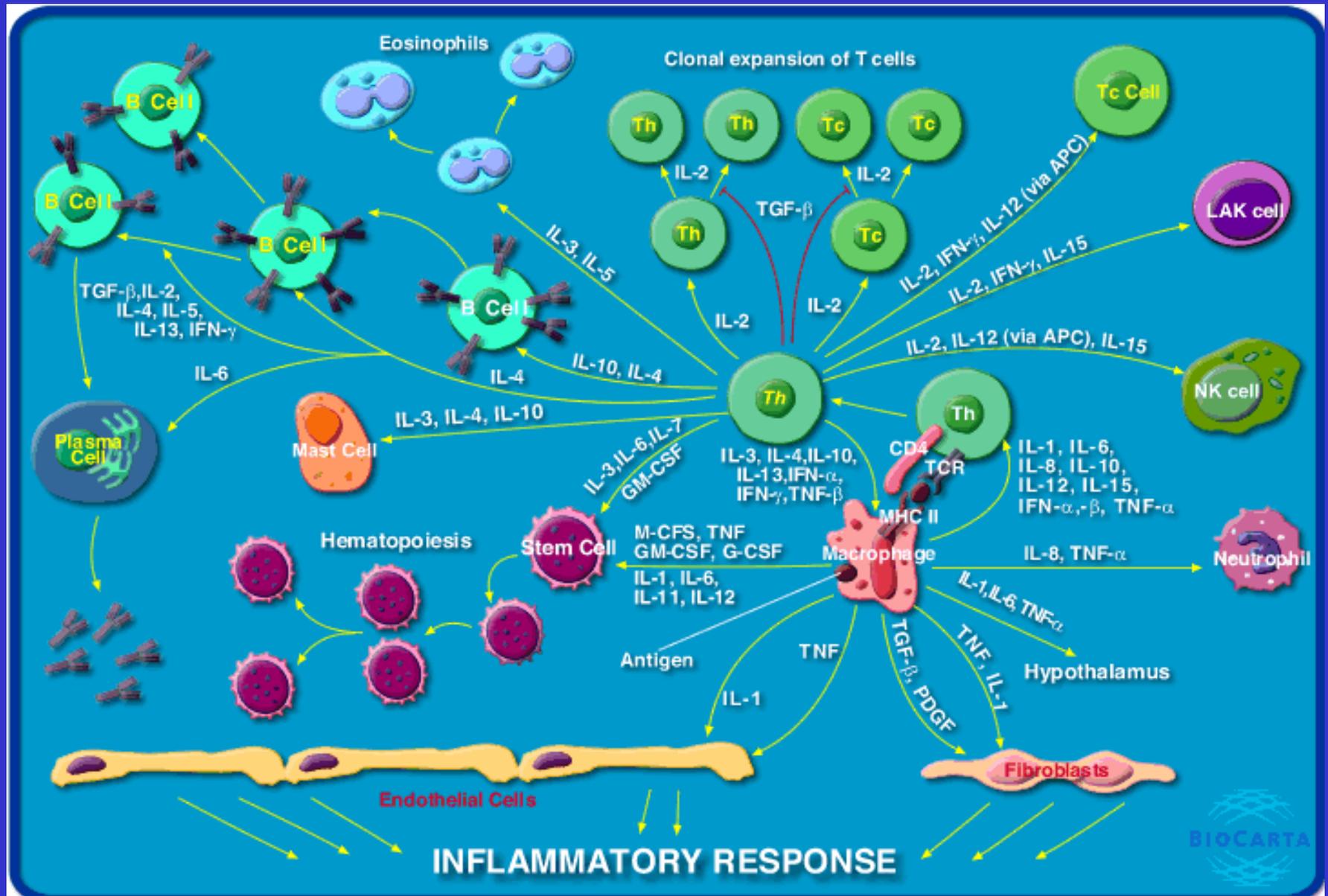
Combination (OR=0.072)

No effect

MS Methionine synthase



Genes in the Immune Pathway



Variants of the *IL4* Pathway

Perturbation of IL-4 activity

‘Components’

Receptor complex

IL4RA, *IL2RA*

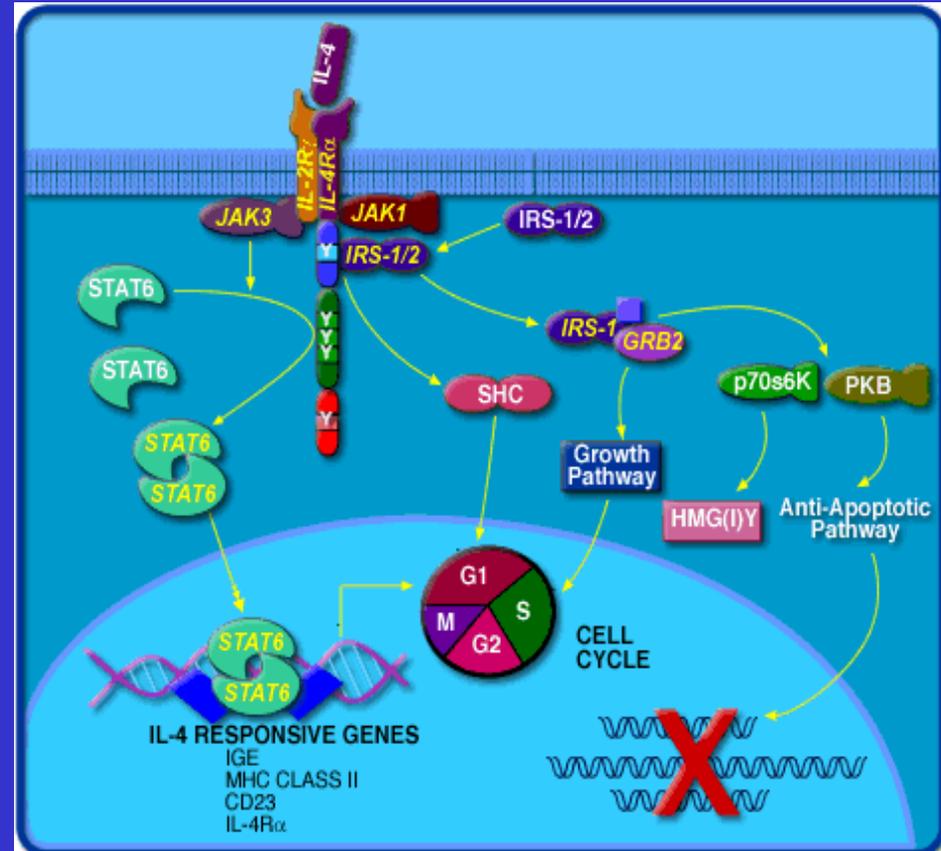
Key genes in signal transduction (*STAT6*, *JAK1*)

‘Close neighbors’

IL5, *IL13* & *RAD50*

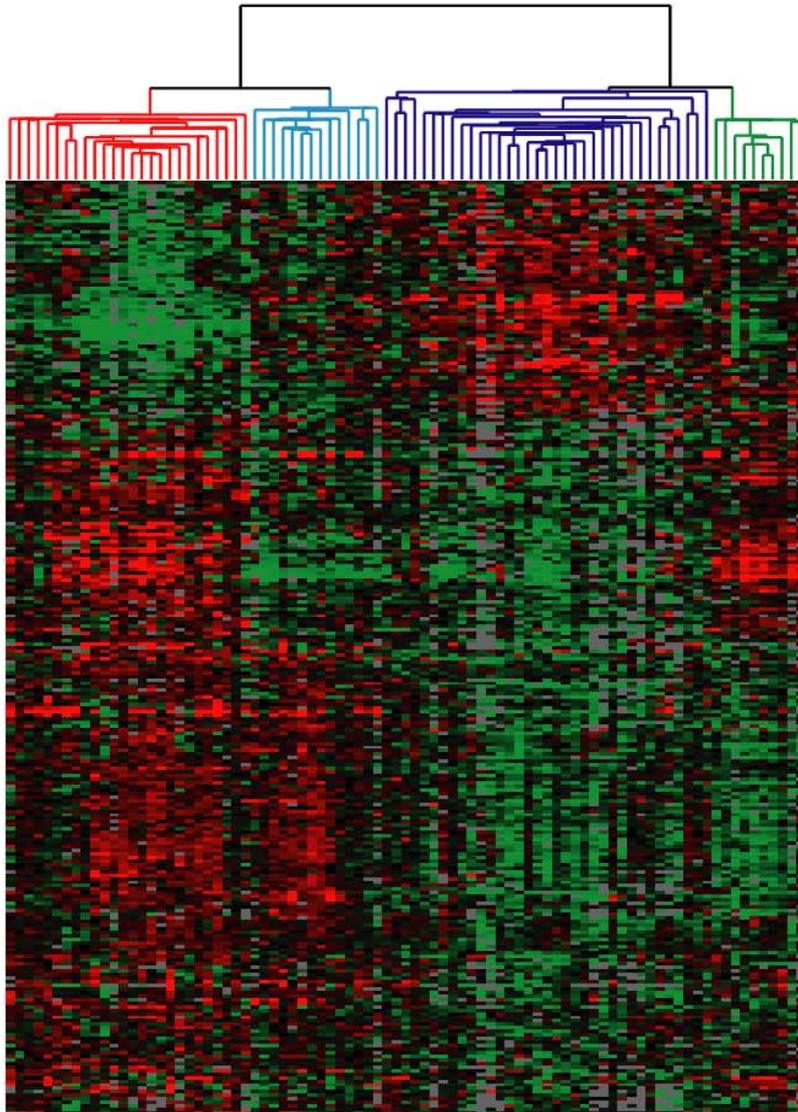


200 kb



**Significance Analysis of Microarrays
to identify genes that correlate
with patient survival**

**SAM264 Gene Set
(254 genes/264 cDNA clones)**



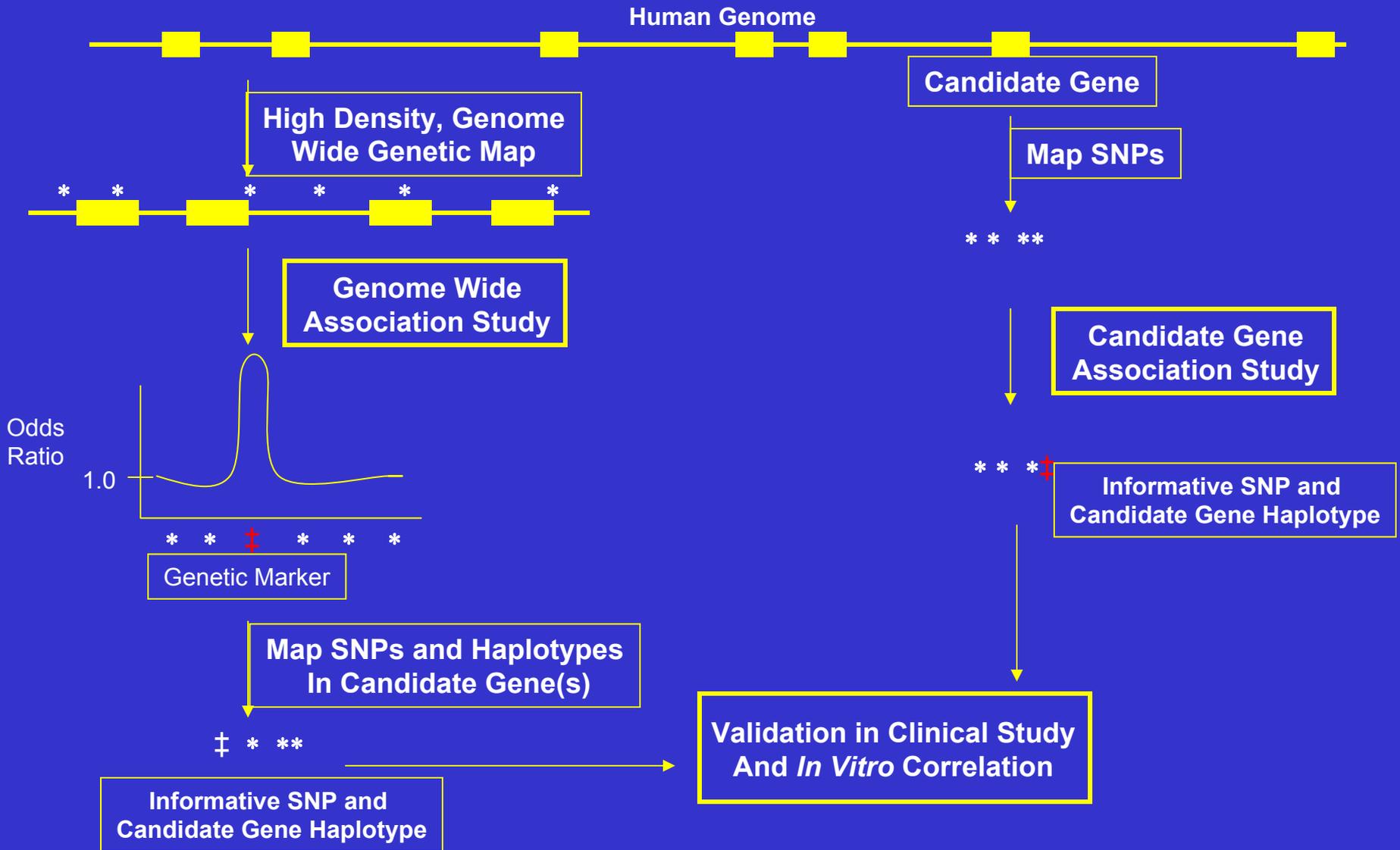
→ **Luminal/ER+ Gene set**

→ **Basal Epithelial Gene set**

→ **Proliferation Gene set**

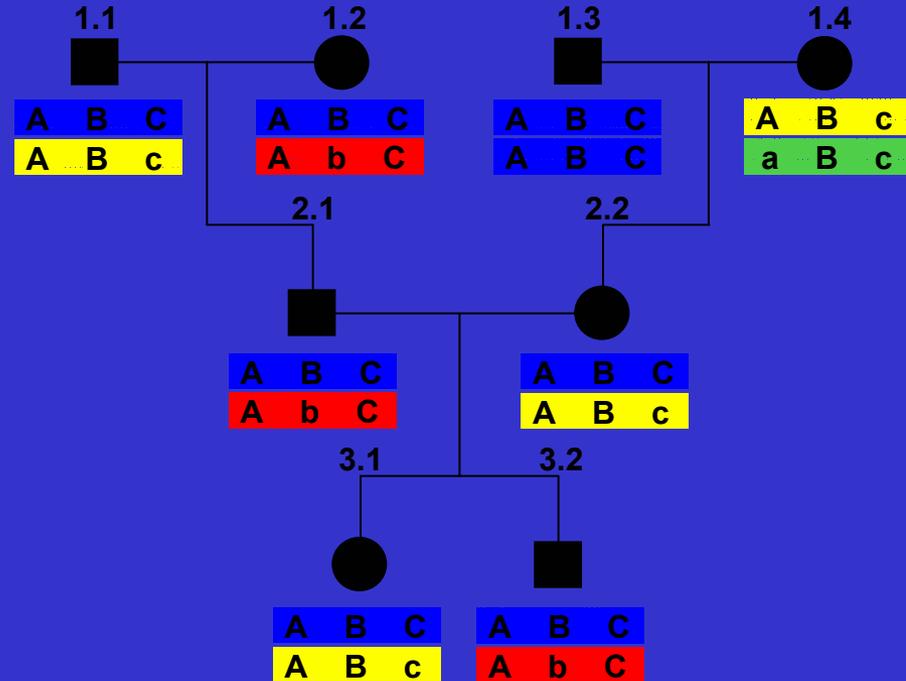


Parallel Approaches To Identifying Genetic Determinants of Disease



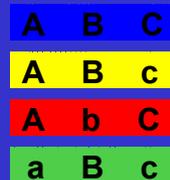
Genotypes and Haplotypes

Individual	Site 1	Site 2	Site 3
1.1	AA	BB	Cc
1.2	AA	Bb	CC
1.3	AA	BB	CC
1.4	Aa	BB	cc
2.1	AA	Bb	CC
2.2	AA	BB	Cc
3.1	AA	BB	Cc
3.2	AA	Bb	CC



Statistical or Laboratory Methods

- Software-PHASE2.0
- Cloning
- Hybrid constructs
- Allele specific amplification

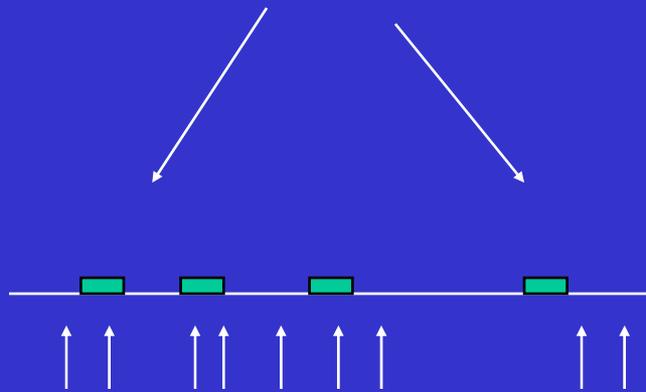


HapMap and Detailed Haplotypes:

Zip Codes vs. Street Numbers



Spaced Markers across genome
High frequency
Define units of genome
“Haplotype blocks”
New goal 1 SNP every 1 kb



Fine mapping of genes
Needed to deconstruct
Haplotypes based on SNPs of
varying frequencies (2-40%)
Find functional components

Estimate of 50,000 to 250,000 Functional SNPs

Utility of The International HapMap

Ability to conduct
indirect studies

“map a locus”

Falling genotype costs

Fixed platforms

Issues in study design

Case-control vs cohort

Definition of haplotypes
across genes of interest

Issues of density
(gene/region specific)

Assay fidelity- genotype
only

Informatic solutions

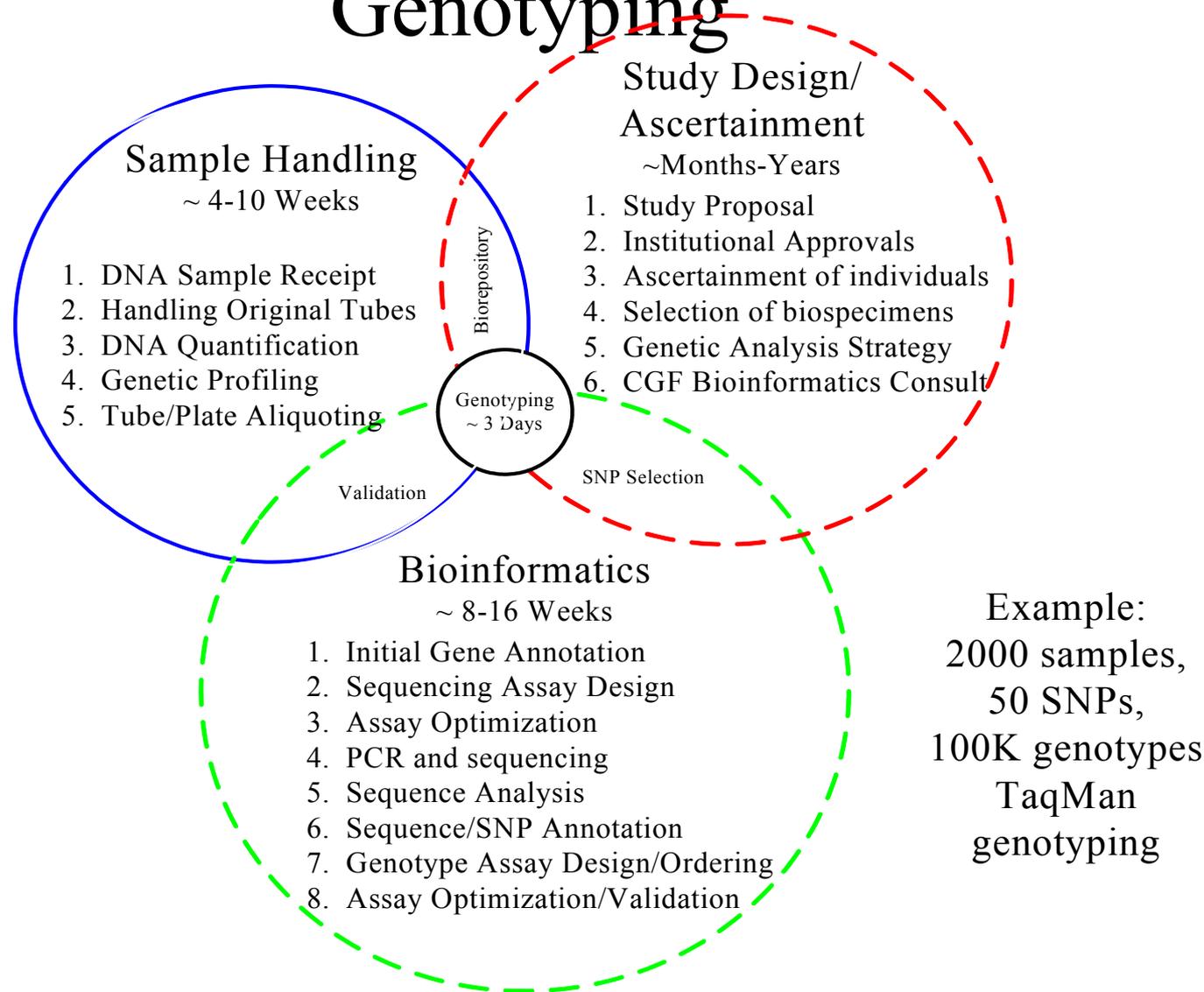
“Navigating the tidal
wave”

<http://hapmap.org>

Choice of Genotype Platforms

Intent	Type	Per day
Extreme Genotyping Chips and Beads	Whole Genome Illumina, Affymetrix, ParAllele, Perlegen	1,000,000
High-Throughput Smaller Chips and Bytes	Pathways and Genes MALDTI-TOF, SNPlex	250,000
Moderate Through-put Probes and Primers	Genes and Haplotypes TaqMan, EPOCH, Pyrosequencing	100,000
Low Through-put RFLP	Gene or Haplotype Gel-based	2,000

Requirements for High Throughput Genotyping

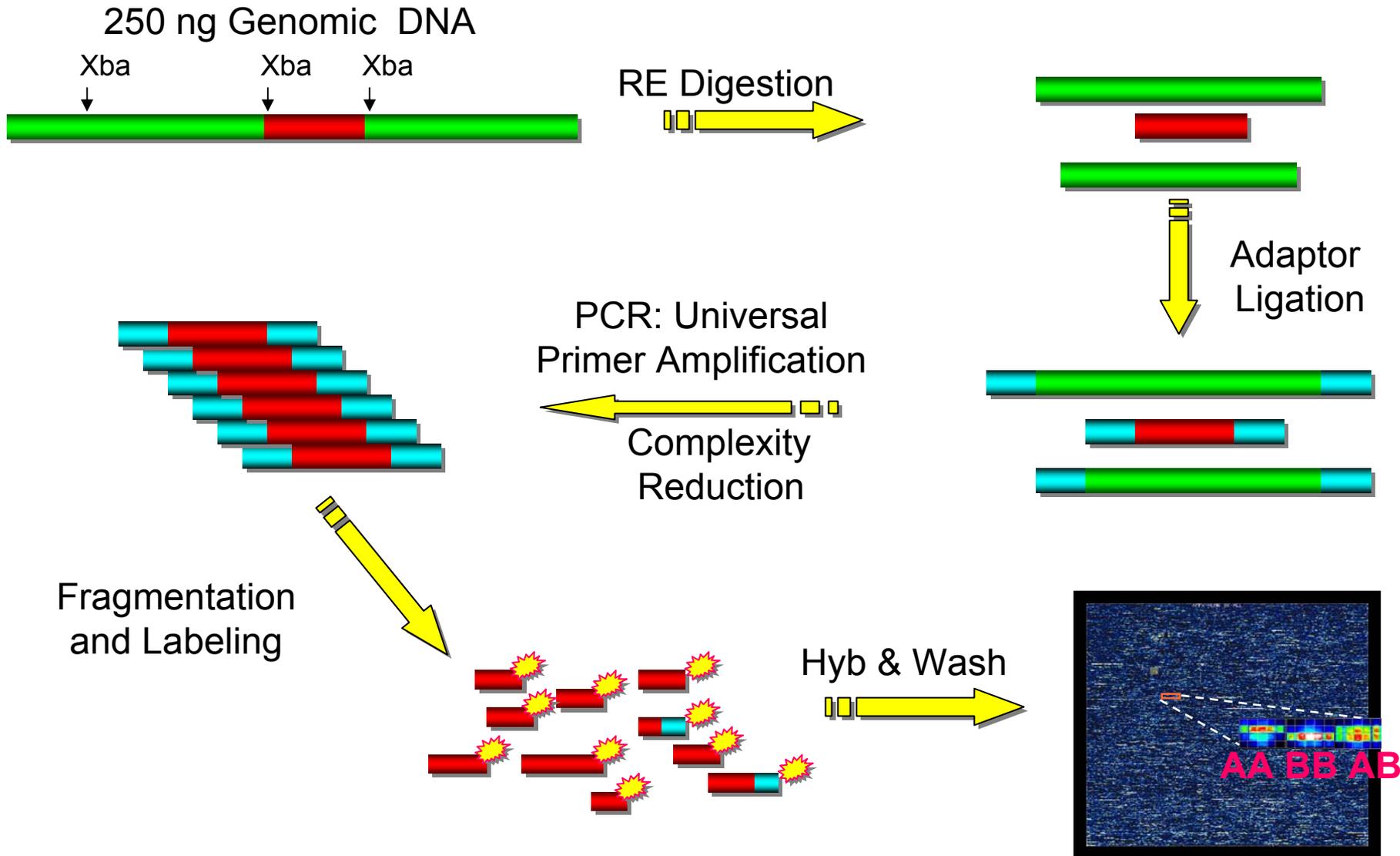


“Extreme Multiplexing”

Approaches:

- Genotype thousands of markers in one reaction!!
- “Simplify genome” followed by amplification and genotyping with hybridization on microarrays
- Allele-specific primer extension followed by ligation and amplification with assortment on microarrays
- Allele-specific gap-fill followed by ligation and amplification and assortment on microarrays

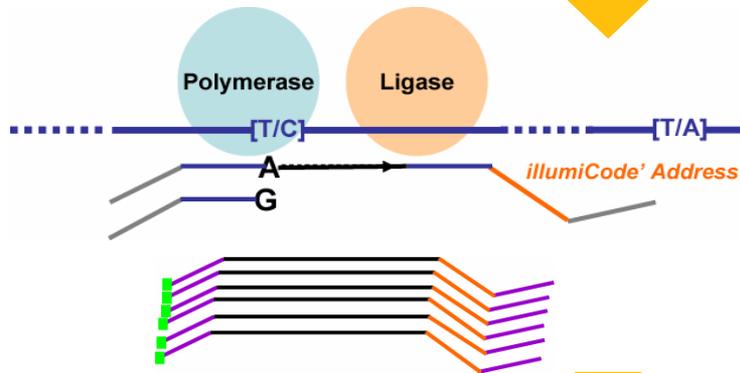
Affymetrix GeneChip[®] Mapping Assay



ILLUMINA GoldenGate™ Assay

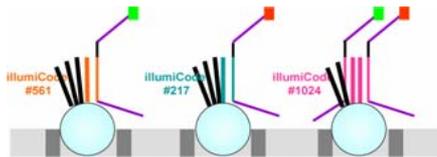


Genomic DNA

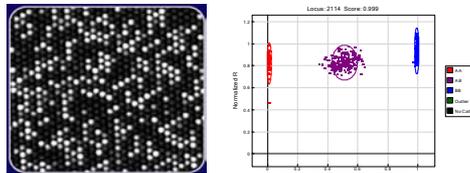


GoldenGate™ Assay
Multiplexed at 1536 loci

Universal Amplification

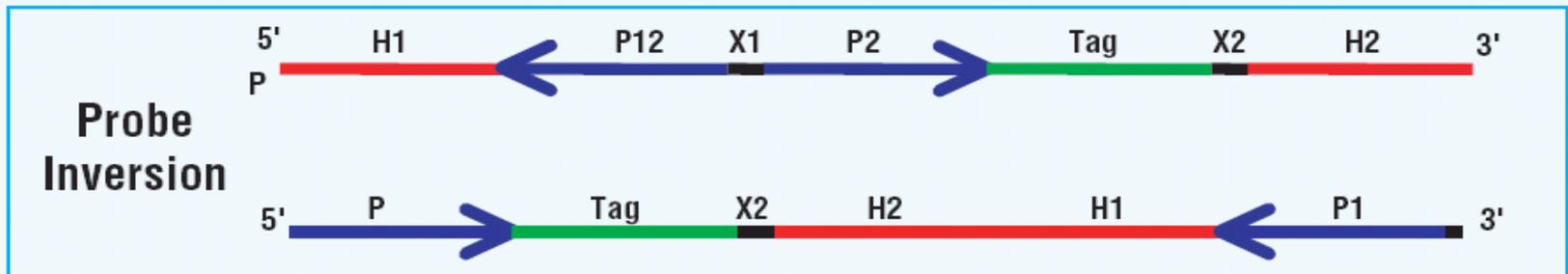
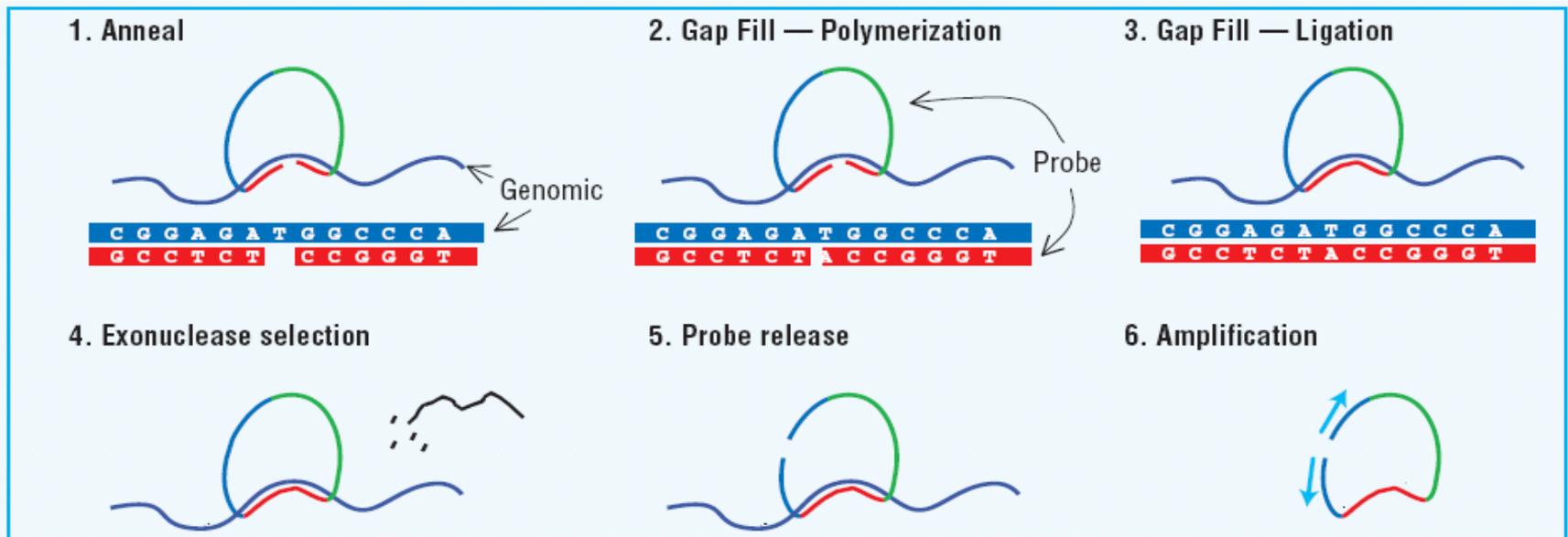


Sentrrix™ BeadArray Hybridization



Automated Array Scan & Analysis
Genotypes with quality score

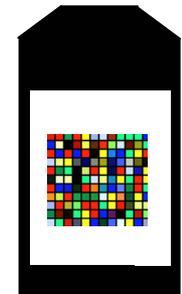
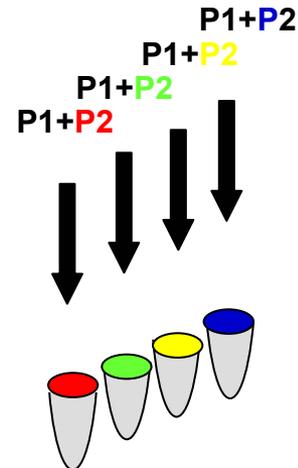
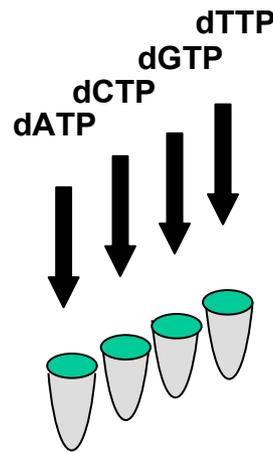
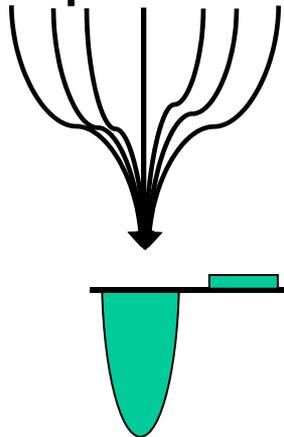
ParAllele Molecular Inversion Probes



Unimolecular interaction enables >12,000 probes to be multiplexed

Multiplex MIP Processing

2,000-10,000
probes



<p>1</p> <p>Mix Probe Pool + Genomic DNA + Enzyme Mix</p> <p><i>Rxn Setup</i></p>	<p>2</p> <p>Divide Mix + Add Nucleotide + Enzyme Mix</p> <p><i>Gap Fill + Ligation</i></p>	<p>3</p> <p>Add Common Primer (Label) + PCR Mix</p> <p><i>Amplification + Labeling</i></p>	<p>4</p> <p>Pool Reactions + Hybridize to chip</p> <p><i>Detection</i></p>
--	---	---	---

Genotyping Throughput

Assays/Rxn	1,536	6,144	12,288
Rxn/Day	960	960	960
Genotypes/ Day	1,474,560	5,988,240	11,796,480
Days/Project	339	84	42

Courtesy of P Kwok

BUT.....

How do you analyze the data?

Example: 1000 cases/controls

250,000 SNPs/samples

500,000,000 SNPs

Informatics

Analysis Platform

Time & FTEs.....

Never leave candidate genes behind....

All studies seek to identify:

Specific variants in unique genes

Whole genome scan

Regional scan

Candidate gene

Effect: Risk for Phenotype

No SNP is sufficient nor necessary....

Accuracy in Molecular Epidemiology

Statistical challenge

Shades of “truth”: $X\%$ vs $(X-12)\%$

Error in genotype

Technical performance

$<0.5\%$ (.2-1.0%) TaqMan, Sequenom, Microsatellites

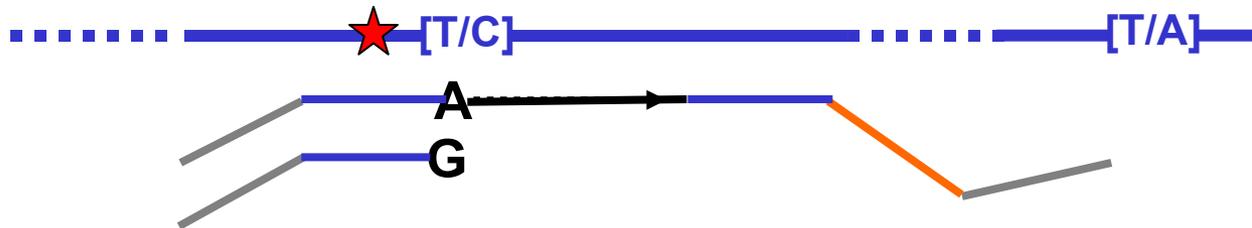
Design

Population-specific adjacent SNPs

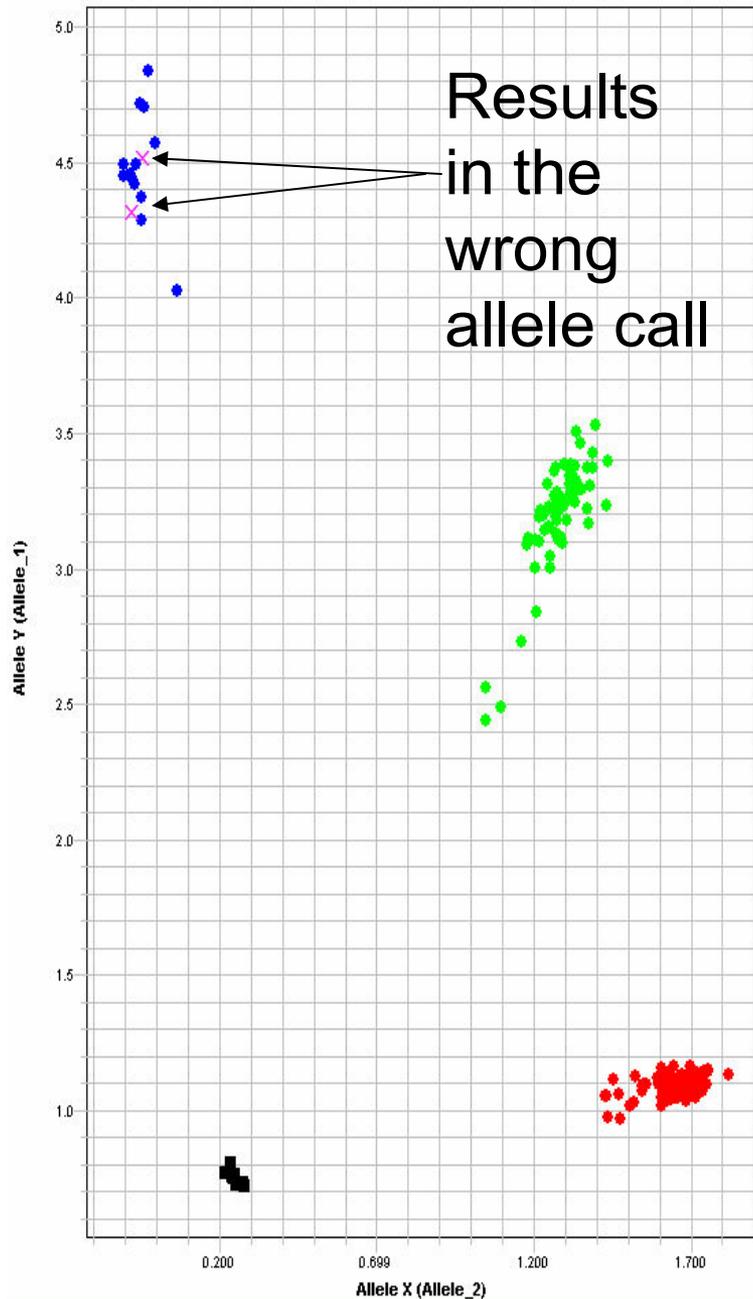
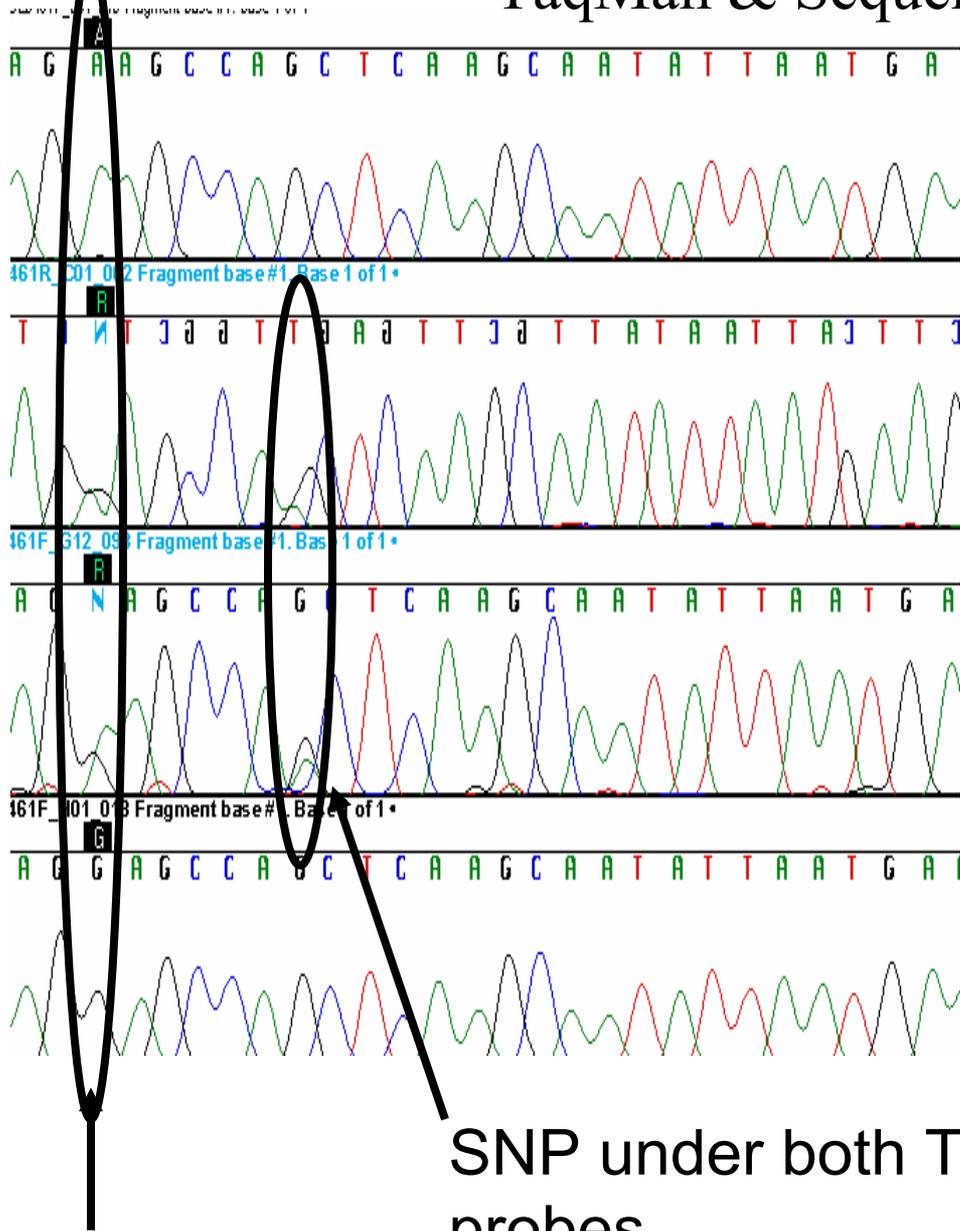
Not a SNP..... (18% monoalleles in SNP500cancer)

Confounding Issues: Neighboring Polymorphisms

- If there are previously unknown polymorphisms in the probe sequences, genotyping errors can occur



TaqMan & Sequencing



SNP500Cancer Cancer Genome Anatomy Project

Number of genes in pipeline: **578**

Total: 7299 SNPs

Genes nominated for molecular epidemiology studies in Core Genotyping Facility of the NCI (<http://cgf.nci.nih.gov>)

102 individuals (Self-Described Ethnicity in Coriell Institute Database)

- **31** Caucasian (test with 4 CEPH parents= 8 +23)
- **24** African Ancestry
- **23** Hispanic
- **24** Pacific Rim

Samples are publicly available from Coriell Institute (<http://locus.umdnj.edu/nigms/>)

Web-posting- daily

All SNPs deposited into db-SNP

Packer et al NAR 2004

NCI > CGAP > SNP500Cancer > Search by SNP

- Welcome, YEAGERM

- Home

- Search by Gene/
Chromosome/
Pathway

- Search by SNP

- Links

- Login Tasks

- Log Out

SNPs matching: **adh1c-02**

dbSNP ID: **rs1693482**

SNP500Cancer ID: ADH1C-02 [dbSNP](#)
 Gene: ADH1C [NCBI map](#)
 Amino acid change: [R272N](#) [Ensembl map](#)
[LocusLink](#)

Sequence of Analyzed Amplicon

```

CTATCTGTTGTTATGGGCTGTAAAAGCAGCTGGAGCAGCCAGAATCATTGCTG
TGGACATCAAYAAGGACAAATTTGCAAAAGGCTAAAAGAGTTGGGTGCCACTGA
ATGCATCAACCCTCAAGACTACAAGAAAACCCATTCAGGAAAGTGCTAAAGGAA
ATGACTGATGGAGGTGTGGATTTTTTCGTTTGAAGTCATCGGTC (A/G) GCTT
GACACCATGGTATGHWCCRTGCATGCCCTGAAAATTTCTGCCTCTGCAACCT
GGAGGATRCATTTAGGCAGYAGAATATACGTATTATGTATAAAGGATATTTT
TAATGATGAATGGAAATTTCCCRTCATCTTTTTTGTACCTGGCTTGTTTAAT
TTA
    
```

Frequency Data (102 anonymized subjects):

Total Completed	Genotypic			Allelic	
	AA	AG	GG	A	G
101	14/101 (0.139)	34/101 (0.337)	53/101 (0.525)	62/202 (0.307)	140/202 (0.693)

[View Subpopulation Frequencies](#)

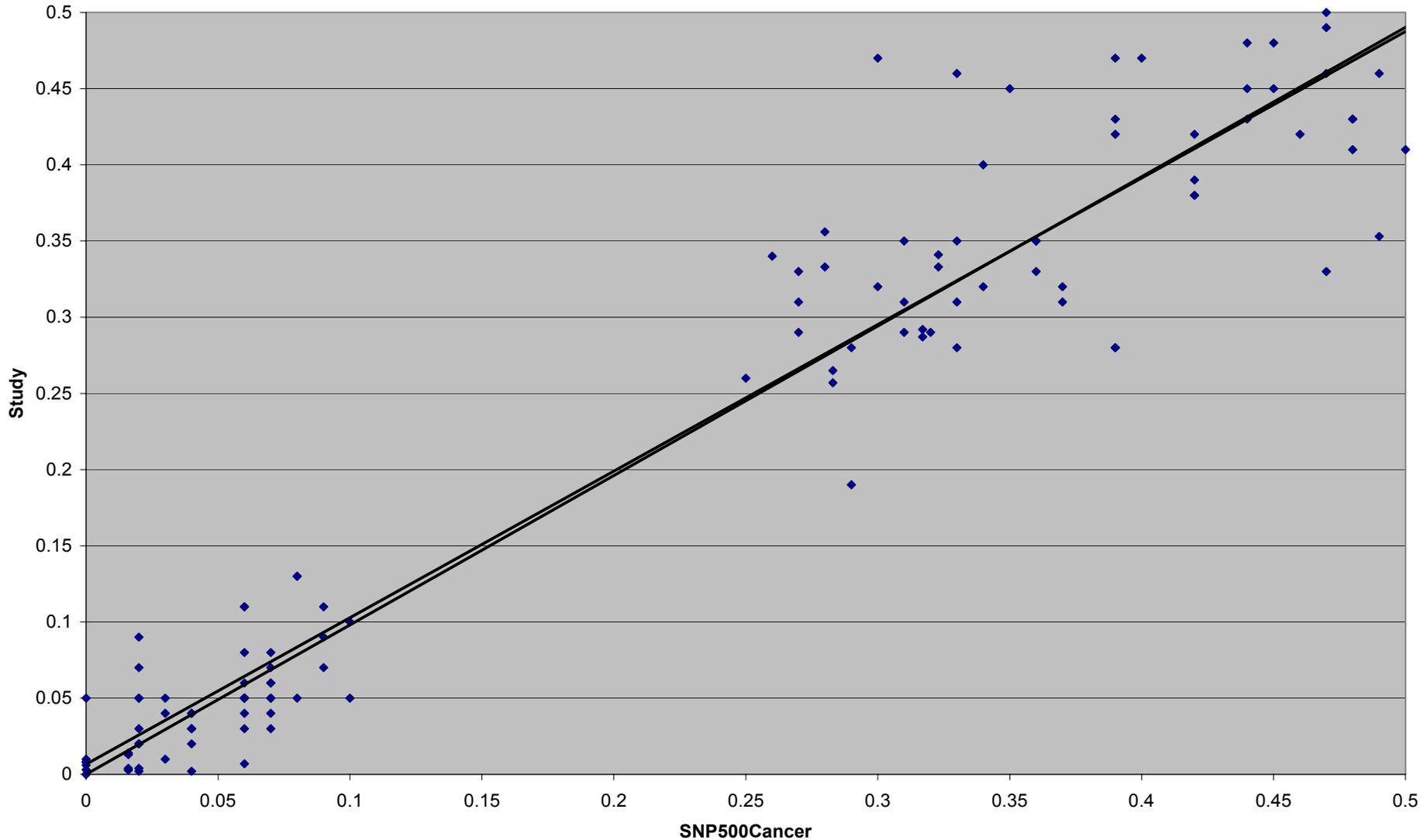
Assays - these frequency results were validated on the following platforms - click to view primers, probes, and conditions:
[Sequencing](#) [TaqMan](#)

Search by SNP

SNP500 estimate vs DCEG study sets (450-2450) for 60 low frequency SNPs ($f < 0.1$)

$y = 0.962x + 0.0066$
 $R^2 = 0.9312$

MAF Estimate < 0.10; > 0.25



($R^2=0.4756$)

($R^2=0.4608$)

New Feature in SNP for Optimized Assays



Cancer Genome Anatomy Project
LuckySNP500 Database

[NCI](#) > [CGAP](#) > [LuckySNP500](#) > [Search by SNP](#)

[Home](#)

[Search by Gene/
Chromosome/
Pathway](#)

[Search by SNP](#)

[Links](#)

[Log In](#)

You requested the additional subpopulation data for **CYP1A2-38**

Frequency Data (280 control samples)

LuckySNP500 ID: CYP1A2-38

Subpopulations	Genotypic			passed <i>HWE?</i>	Allelic	
	CC	CT	TT		C	T
Total Completed	266/280 (0.950)	13/280 (0.046)	1/280 (0.004)	-	545/560 (0.973)	15/560 (0.027)
Afr/Afr American	76/76 (1.000)	0/76 (0.000)	0/76 (0.000)	passed	152/152 (1.000)	0/152 (0.000)
Caucasian	66/66 (1.000)	0/66 (0.000)	0/66 (0.000)	passed	132/132 (1.000)	0/132 (0.000)
Hispanic	46/49 (0.939)	3/49 (0.061)	0/49 (0.000)	passed	95/98 (0.969)	3/98 (0.031)
Pacific Rim	78/89 (0.876)	10/89 (0.112)	1/89 (0.011)	passed	166/178 (0.933)	12/178 (0.067)

[Explanation of subpopulations](#)

Search by SNP

- by SNP identifier**
to display SNP details, allele and genotype frequencies

Enter the dbSNP ID or internal SNP ID:

[search hints](#)

[disclaimer](#)




Cancer Genome Anatomy Project
 SNP500Cancer Database

NCI > CGAP > SNP500Cancer > Search by SNP

- Welcome, YEAGERM
- Home
- Search by Gene/
Chromosome/
Pathway
- Search by SNP
- Links
- Login Tasks
- Log Out

You requested the subpopulation data for **rs1693482** (ADH1C-02):

Frequency Data (102 anonymized subjects)

dbSNP ID: rs1693482						
Subpopulations	Genotypic			passed <i>HWE?</i>	Allelic	
	AA	AG	GG		A	G
Total Completed	14/101 (0.139)	34/101 (0.337)	53/101 (0.525)	-	62/202 (0.307)	140/202 (0.693)
Afr/Afr American	0/24 (0.000)	3/24 (0.125)	21/24 (0.875)	passed	3/48 (0.063)	45/48 (0.938)
Caucasian	7/31 (0.226)	16/31 (0.516)	8/31 (0.258)	passed	30/62 (0.484)	32/62 (0.516)
Hispanic	4/23 (0.174)	10/23 (0.435)	9/23 (0.391)	passed	18/46 (0.391)	28/46 (0.609)
Pacific Rim	3/23 (0.130)	5/23 (0.217)	15/23 (0.652)	passed	11/46 (0.239)	35/46 (0.761)

Explanation of subpopulations $F_{st} = 0.164$

Search by SNP

- [by SNP identifier](#) [search hints](#)

Enter the dbSNP ID or internal SNP ID:

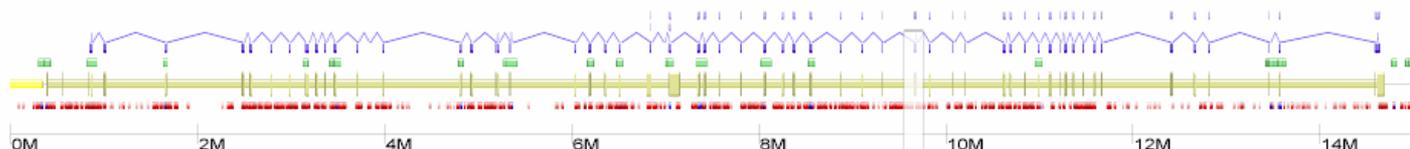
ATM

THEANNOTATOR [logout](#)

SNP

Gene

Search



ATCAGGGTAAACTAAATTAGTAGTCAAAGTTAACTACTAATAGTCTACTGTTGACCAG
 AAGCCCTACCGATGATATAACAACACAATTTTGTATATGTATTATATACTGTATTCTTA
 AATAAAGTAGCTTGAGAAAAGAAAATG **T** AAATGATAAGGAAGAGAAAATGCA
 TTTACAGAACTGTA CTGTGTTTATCAA **-** TTTATATCATCTGTTTACAAAGAA
 TCATCTGTCTGAAATTGTAGGCAACCACTGCTGGTTGATATGTA CTTTAGGCCTCAATCT
 ACAGAA TGATCAAACAATTACCTTTTTCTTGTAATGGCATGACTTTTCTCTGCATCTT
 GGAAGCACCTCCAGCCACTGTGGTGTAAATCAAGGCTTATAGTATTGCAGTAAACACAGT
 GAAAAAATACATGAGAACCTGGAGAGATTACTTTTCATTTCTGGATACAATTTACTAGAG
G
 AAATGAACTGCTCATGTGGAATTATTAGTGTCACATGGCATTTTAAGCAGATACTCACA

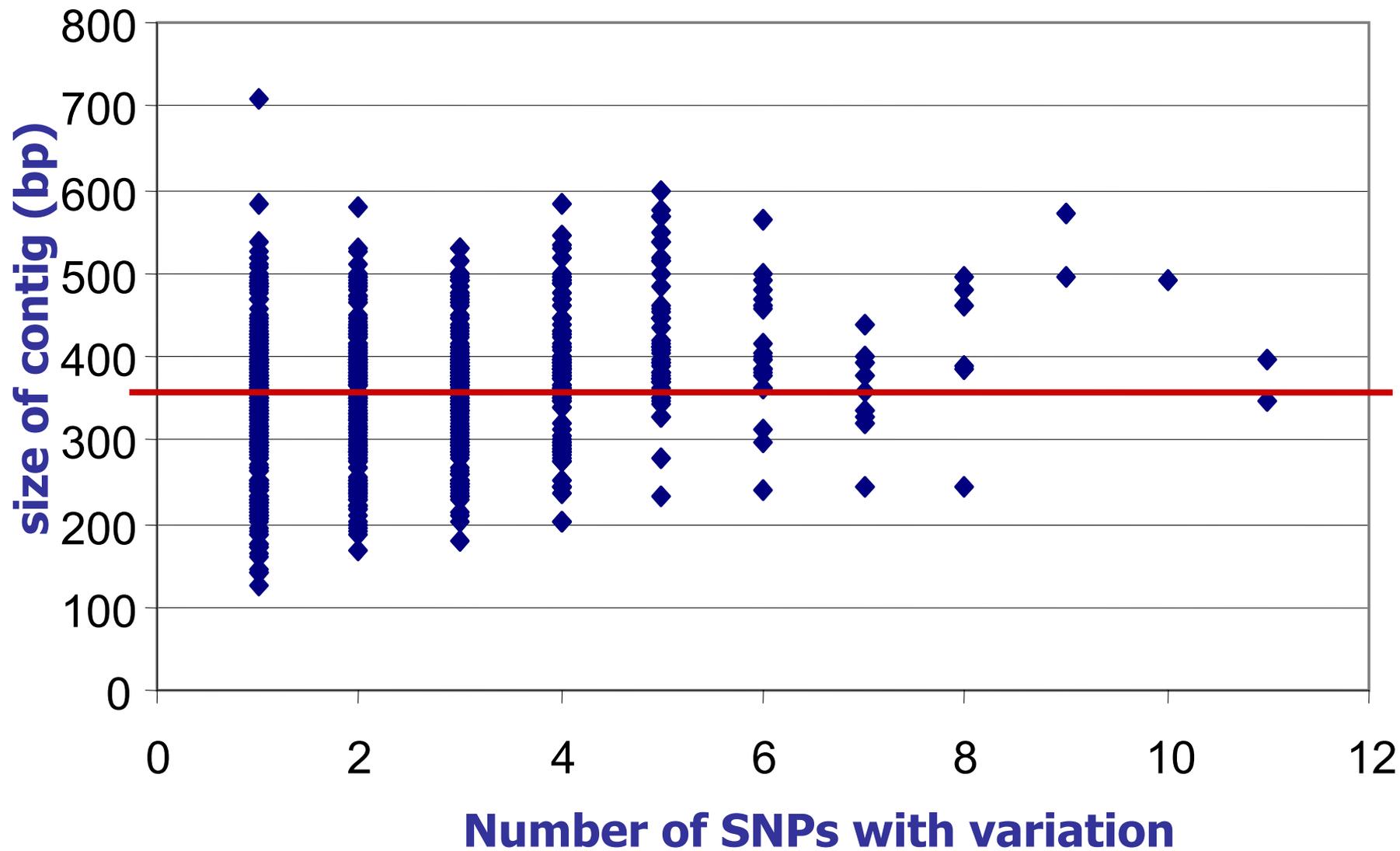
rs620613 X

Avg. Het. 0.49875

C/T

info Submit

Distribution of # of SNPs per contig



SNP500 Cancer

578 Genes in pipeline (~11 SNPs per gene)

92% Contigs sequenced with at least 1 SNP with variation

5901 Completely Analyzed

4890 (83%) with variation

1011 (17%) NO variation (in 204 chromosomes)

357 bp Average size of contigs (median 366bp)

2.14 SNPs per 357bp was not anticipated!!

1 Probable SNP every 154 bp (comparable to Wong et al *Genome*

Res 2003)



YEAGERM

[Home](#)

[Search by Gene/
Chromosome/
Pathway](#)

[Search by SNP](#)

[Links](#)

[Login Tasks](#)

[Log Out](#)

Assay number 539: 102 individuals completed

Strand: +

	Concentration	5' dye	Sequence	3' dye	Allele
Probe 1	150 nM	FAM	ATCGGTCGGCTTGA	MGB	G
Probe 2	400 nM	VIC	CATCGGTCAGCTTG	MGB	A
Primer F	900 nM		TGACTGATGGAGGTGTGGATTTT		
Primer R	900 nM		GCAGAAATTCAGGGCATGTC		

Annealing temperature: 63 degrees centigrade

Method: 10 ng of lyophilized sample DNA was used to do a 5ul Taqman reaction. The reactions were done in a 384 (96*4) well plate format. Four controls (Coriell DNA controls) for each genotype as well as NTCs (no template controls) were put on the plate along with the samples. 2.5 ul of the 2X Universal Mater Mix (ABI) and assay-specific concentrations of primers and probes were used in the reaction. Assay-specific thermo-cycling conditions were used:

- Step 1 -- 50°C - 2 Minutes - AmpErase UNG activation (ABI)
- Step 2 -- 95°C - 10 Minutes - Enzyme activation
- Step 3 -- 0:30 Seconds - Template Denaturation
 - 92°C if using **3'MGB** quencher
 - 95°C if using **3'TAMRA** quencher
- Step 4 -- 60°C - 1:00 Minute - Annealing (assay-specific)
- Step 5 -- GO TO 3 49 TIMES
- Step 6 -- 4°C - Hold
- Step 7 -- END

The plate was then read on the ABI 7900HT sequence detection system. The SDS software graphs the results of allelic discrimination run on a scatter plot of Allele 1 Rn versus Allele 2 Rn. The software represents each well of the 384-well plate as a spot on the graph. The genotypic segregation is displayed in the allelic plot. The plot contains four distinct clusters, which represent the NTCs (no template controls) and three possible genotypes that cluster along the horizontal, vertical and diagonal axes and represent the Allele 1, Allele 2 and Allele 1/Allele 2 respectively. This variation is due to differences in the extent of PCR amplification. The data are then exported in text format for further analysis.

The Cohort Consortium for Breast and Prostate Cancers

- Fine sequence analysis
- Haplotype determination
 - Haplotype tagging SNPs
 - Capture >95% of common variation
 - No is Firm
 - Yes means we need to look closely at functional SNPs
- Genotype analysis of ht-SNPs across all studies
 - Pooled analysis for main effects

Re-sequence Analysis in Breast & Prostate Cancer

- Cohort Consortium for Breast and Prostate
 - MIT/USC, Harvard, EPIC, CEPH, NCI
 - ~8,000 cases and ~8,000 controls for each
- Re-sequence in Cases (n=190)
- Haplotype determination (n=760)
 - CEPH trios & Unrelated MEC populations
- 53 genes
 - IGF pathway & Sex steroid metabolism
- Pooled Association Studies

Re-sequence Analysis in Breast & Prostate Cancer

STEP 1

Resequencing analysis of

- All exons

- 5' and 3' UTR

- ECR regions (>80% homology between human and mouse)

Bio-informatic identification of SNPs every 1-2 kb in introns

STEP 2

Analyze SNPs with > 5-10% minor allele frequency in

- ~350 unrelated (5 ethnic groups)

- > 140 trios

STEP 3

Determine ht-SNPs for large studies

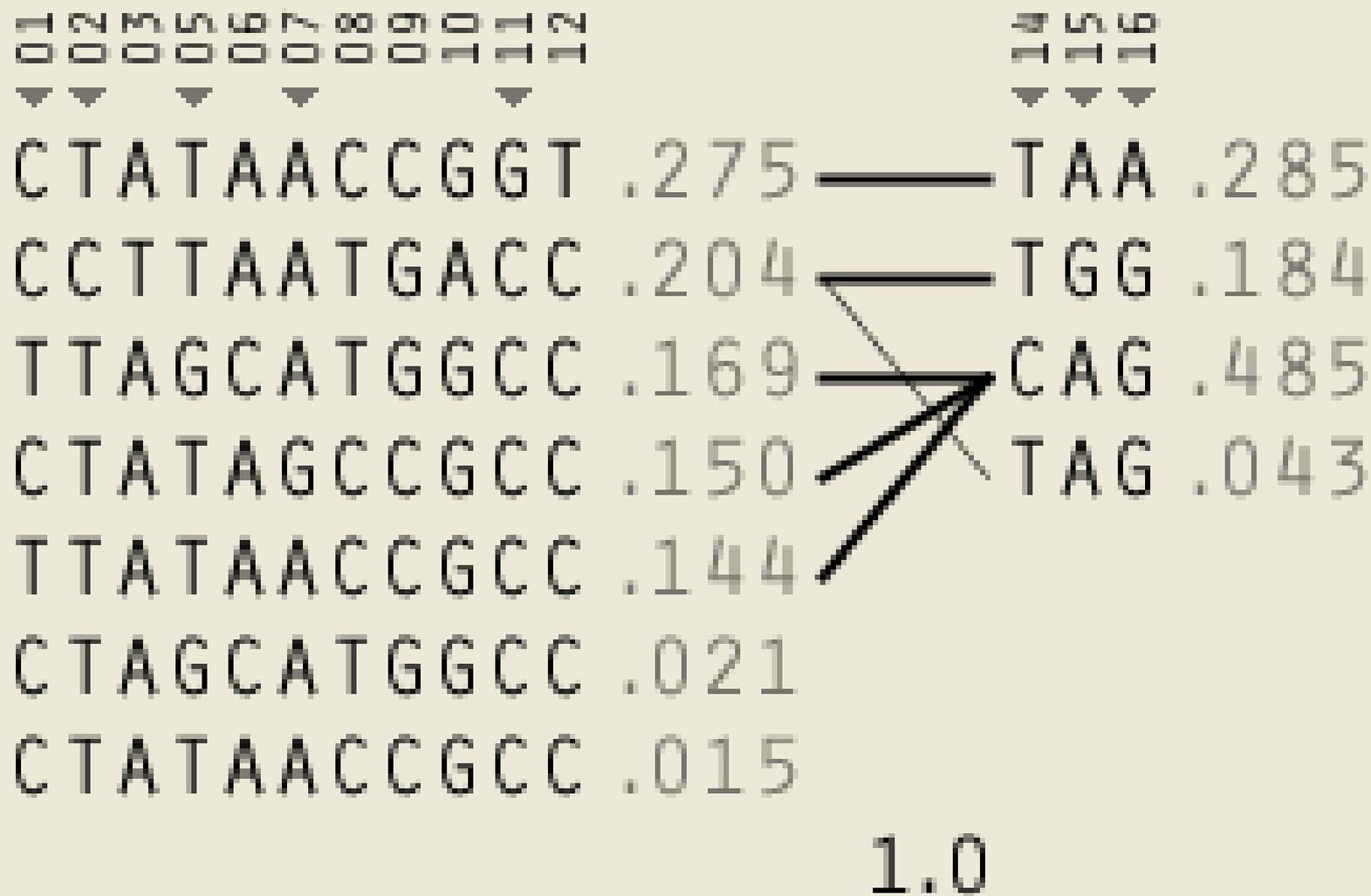
FY04 Genes in the pipeline: from sequence to haplotype to analysis

- Full analysis underway
 - AR*
 - HD17B1*
 - CYP19*
- Assays in pipeline
 - CYP17*
 - PGR*
 - ESR2*
- ht-SNPs Ready
 - GHR*
 - HSD17B4*
 - CYP1B1*

CYP1B1-Analysis

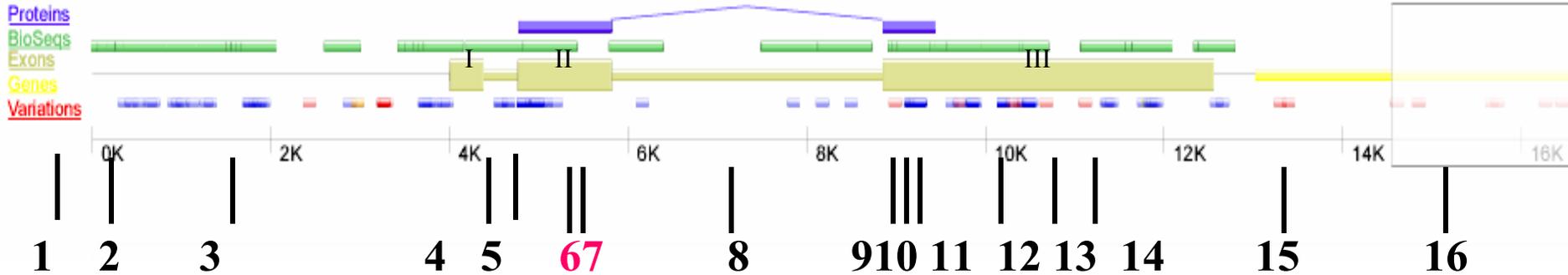
- Sequenced in SNP500 (n=102 from 4 ethnic groups) and 92 DNAs (5 ethnic groups of MEC)
- 64 SNPs found, 37 with Minor Allele Frequency >10%
- 16 ht SNPs genotyped in CEPH ethnic panel and CEPH trios
 - High failure rate 14/30 (*odd gene....*)

ht-SNPs for CYP1B1 in European Caucasians

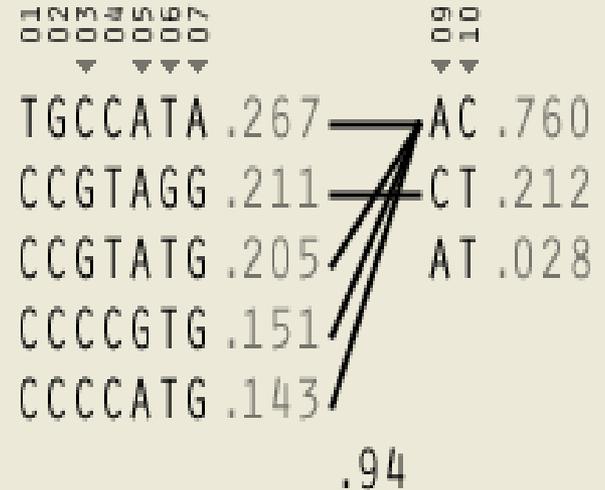
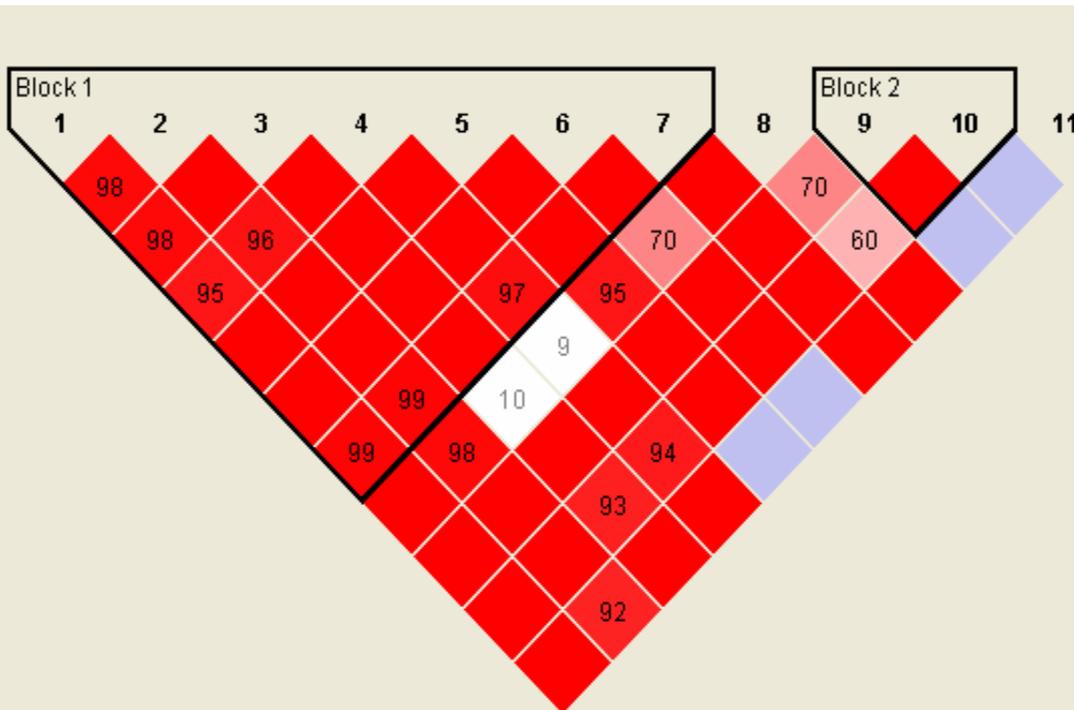


8 htSNPs Captures 98.6% Diversity

positions of 16 SNPs in *CYP1B1*



140 unrelated CEPH trio parents



Central questions

- Scope of candidate gene analysis
 - Pathways- classical vs array driven
 - ht-SNPs
 - Screen vs across all studies
- Determine by platform
 - Cost and availability
- Centralized vs Separate institutions
 - QC vs cost vs efficiency
- Whole genome amplified DNA
- Main effect plus ? Co-variates

Acknowledgements

CGF-NCI

M Yeager

B Welch

A Crenshaw

A Bergen

NCI

J Fraumeni

N Rothman

CGAP-NCI

L Burdett

B Strausberg

B Packer

D Gerhardt

S Presswalla