Overview of SNP Genotyping

Debbie Nickerson

Department of Genome Sciences
University of Washington
debnick@u.washington.edu

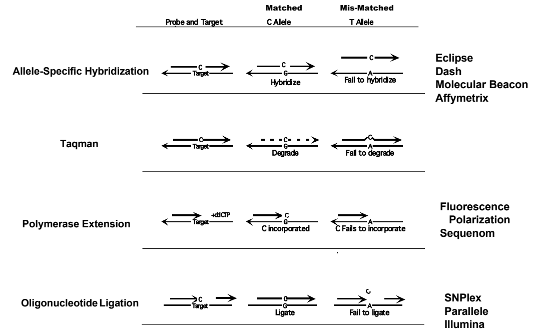
SNP Genotyping - Overview

- Project Rationale
- Genotyping Strategies/Technical Leaps
- Data Management/Quality Control

SNP Project Rationale

- Heritability
- Power - Number of Individuals
- Number of SNPs - Candidate Gene, Pathway, Genome
  5-10 SNPs, 400 to 1,000, 10K, 500K
- DNA requirements
- Cost

SNP Genotyping



SNP Typing Formats

Scale

- Microtiter Plates - Fluorescence — Low
  eg. Taqman - Good for a few markers - lots of samples - PCR prior to genotyping
- Size Analysis by Electrophoresis — Medium
  eg. SNPlex - Intermediate Multiplexing reduces costs - Genotype directly on genomic DNA - new paradigm for high throughput
- Arrays - Custom or Universal — High
  eg. Illumina, ParAllele, Affymetrics - Highly multiplexed - 96, 1,500 SNPs and beyond (500K+)

Defining the scale of the genotyping project is key to selecting an approach:

	1000 individuals
5 to 10 SNPs in a candidate gene - Many approaches (expensive ~ 0.60 per SNP/genotype)	$6,000
48 (to 96) SNPs in a handful of candidate genes (~ 0.25 to 0.30 per SNP/genotype)	$~29,000
384 - 1,536 SNPs - cost reductions based on scale (~0.08 - 0.15 per SNP/genotype)	$57,600-122,880
300,000 to 500,000 SNPs defined format (~0.002 per SNP/ genotype)	$800,000
10,000-20,000 SNPs - defined and custom formats (~0.03 per SNP/genotype)	$>250,000

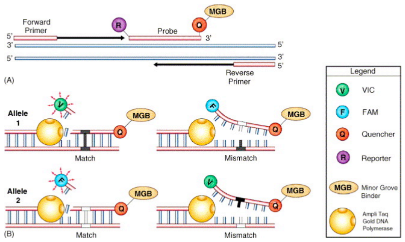
Many Approaches to Genotype a Handful of SNPs

PCR region prior to SNP genotyping - Adds to cost
- Many use modified primers - the more modified, the higher the cost

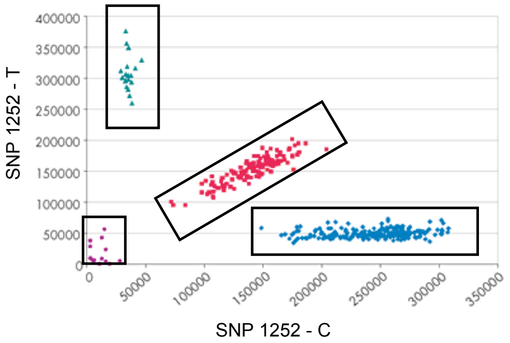
- **Taqman**
- **Single base extension - Fluorescence Polarization**
  Sequenom - Mass Spec
- **Eclipse**
- **Dash**
- **Molecular Beacons**

Taqman

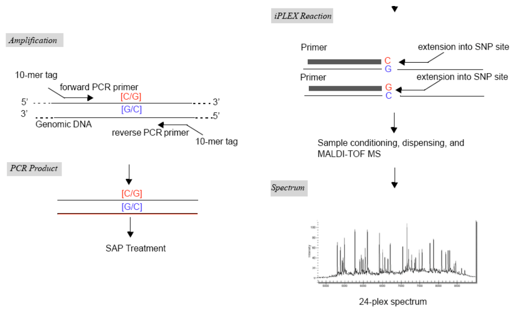
Genotyping with fluorescence-based homogenous assays (single-tube assay) = 1 SNP/ tube



Genotype Calling - Cluster Analysis



Genotyping by Mass Spectrometry - 24



Technological Leap - No advance PCR

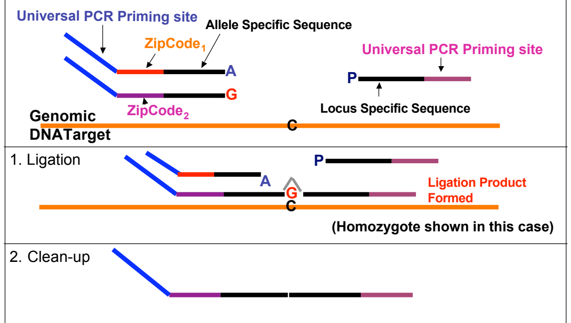
Universal PCR after preparing multiple regions for analysis -

Several based on primer specific on genomic DNA followed by PCR of the ligated products - different strategies and different readouts.

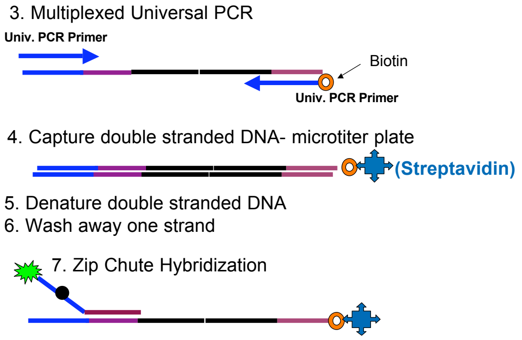
SNPlex, Illumina, Parallele (Affymetrix)

Also, reduced representation - Affymetrix
- cut with restriction enzyme, then ligate linkers and amplify from linkers and follow by chip hybridization to read out.

SNPlex Assay - 48 SNPs



PCR & ZipChute Hybridization



3. Multiplexed Universal PCR

4. Capture double stranded DNA- microtiter plate

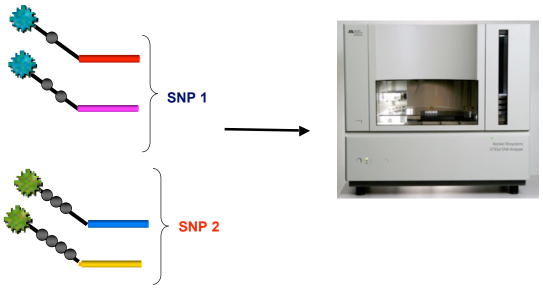
5. Denature double stranded DNA

6. Wash away one strand

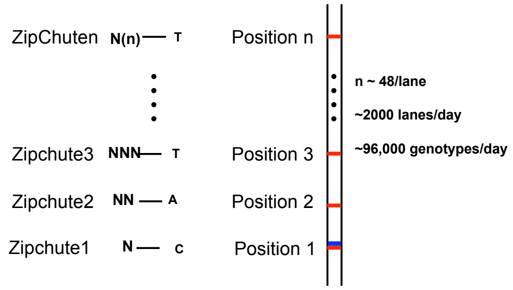
7. Zip Chute Hybridization

Detection

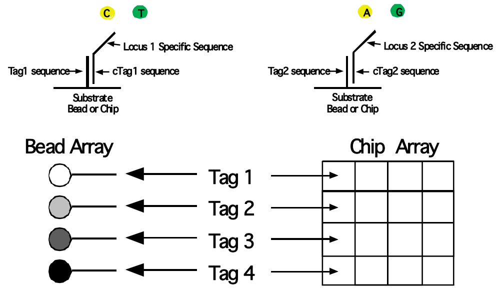
9. Characterize on Capillary Sequencer



SNPlex Readout



Multiplexed Genotyping - Universal Tag Readouts



Illumina

Multiplex ~96 - 20,000 SNPs

Not dependent on primary PCR

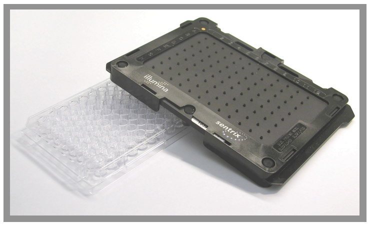
ParAllele

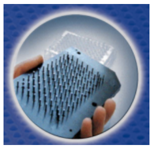
Affymetrix

Arrays - High Density Genotyping Thousands of SNPs and Beyond

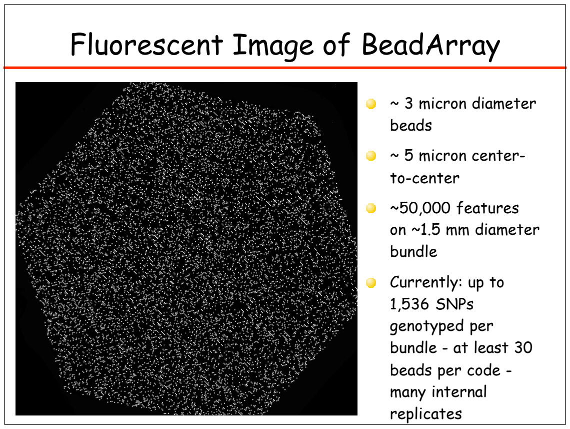
- “Bead” Arrays - Illumina
  - Manufactured by self-assembly
  - Beads identified by decoding

Sentrix™ Platform



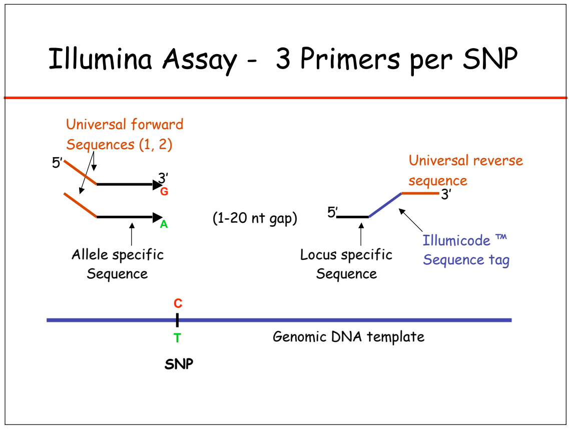


- Sentrix™ 96 Multi-array Matrix matches standard microtiter plates (96 - 1536 SNPs/well)
- Up to ~140,000 assays per matrix

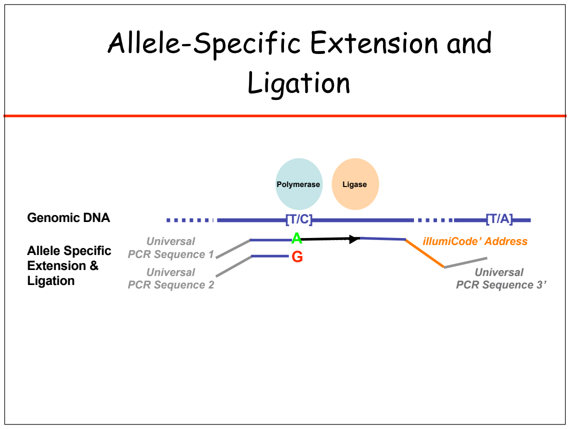


Fluorescent Image of BeadArray

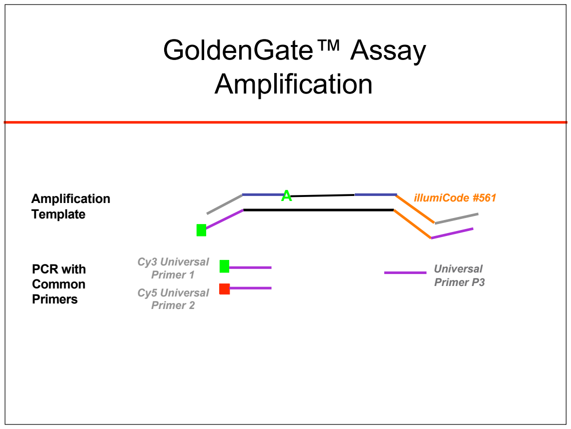
- ~ 3 micron diameter beads
- ~ 5 micron center-to-center
- ~50,000 features on ~1.5 mm diameter bundle
- Currently: up to 1,536 SNPs genotyped per bundle - at least 30 beads per code - many internal replicates



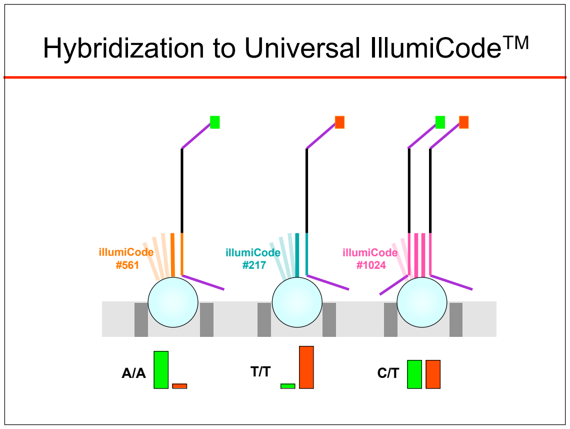
Illumina Assay - 3 Primers per SNP



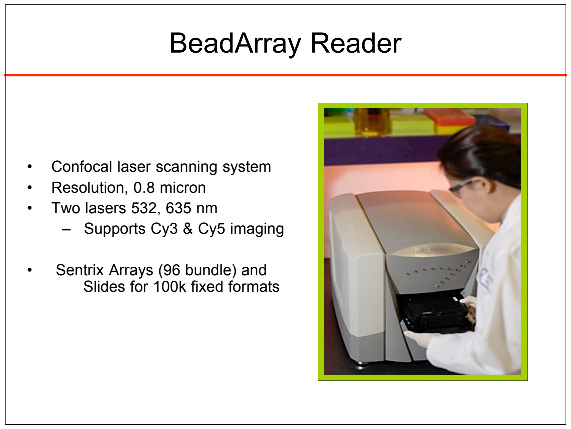
Allele-Specific Extension and Ligation



GoldenGate™ Assay Amplification

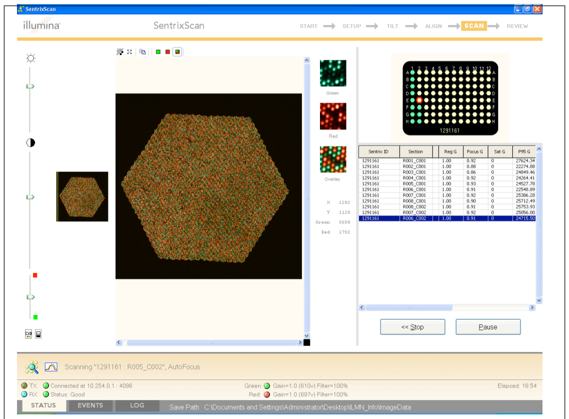


Hybridization to Universal IllumiCode™



BeadArray Reader

- Confocal laser scanning system
- Resolution, 0.8 micron
- Two lasers 532, 635 nm
  - Supports Cy3 & Cy5 imaging
- Sentrix Arrays (96 bundle) and Slides for 100k fixed formats



Sentrix ID	Section	Reg G	Focus G	Sat G	P95 G
1291161	R001_C001	1.00	0.92	0	27624.34
1291161	R002_C001	1.00	0.88	0	22274.88
1291161	R003_C001	1.00	0.86	0	24849.46
1291161	R004_C001	1.00	0.92	0	24264.41
1291161	R005_C001	1.00	0.93	0	24527.78
1291161	R006_C001	1.00	0.91	0	22548.89
1291161	R007_C001	1.00	0.92	0	25386.28
1291161	R008_C001	1.00	0.90	0	25712.49
1291161	R008_C002	1.00	0.91	0	25753.93
1291161	R007_C002	1.00	0.92	0	25056.00
1291161	R006_C002	1.00	0.91	0	24715.50

X 1282
Y 1128
Green 5608
Red 1702

<< Stop
Pause

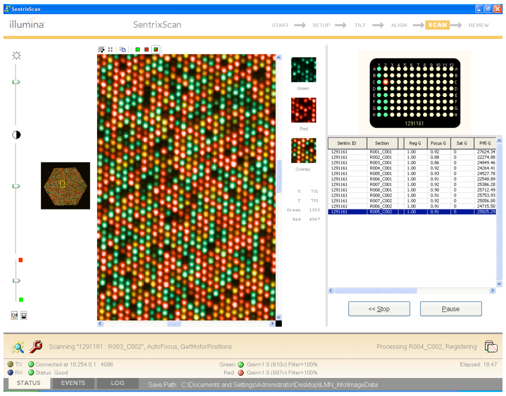
Scanning "1291161 : R005_C002", AutoFocus

TX ● Connected at 10.254.0.1 : 4096
RX ● Status: Good

Green: ● Gain=1.0 (610v) Filter=100%
Red: ● Gain=1.0 (697v) Filter=100%

Elapsed: 16:54

STATUS EVENTS LOG Save Path: C:\Documents and Settings\Administrator\Desktop\ILMN_Info\ImageData



Sentrix ID	Section	Reg G	Focus G	Sat G	P95 G
1291161	R001_C001	1.00	0.92	0	27624.34
1291161	R002_C001	1.00	0.88	0	22274.88
1291161	R003_C001	1.00	0.86	0	24849.46
1291161	R004_C001	1.00	0.92	0	24264.41
1291161	R005_C001	1.00	0.93	0	24527.78
1291161	R006_C001	1.00	0.91	0	22548.89
1291161	R007_C001	1.00	0.92	0	25386.28
1291161	R008_C001	1.00	0.90	0	25712.49
1291161	R008_C002	1.00	0.91	0	25753.93
1291161	R007_C002	1.00	0.92	0	25056.00
1291161	R006_C002	1.00	0.91	0	24715.50
1291161	R005_C002	1.00	0.91	0	25825.29

X 731
Y 753
Green 1393
Red 4967

<< Stop
Pause

Scanning "1291161 : R003_C002", AutoFocus, GetMotorPositions

Processing R004_C002, Registering

TX ● Connected at 10.254.0.1 : 4096
RX ● Status: Good

Green: ● Gain=1.0 (610v) Filter=100%
Red: ● Gain=1.0 (697v) Filter=100%

Elapsed: 18:47

STATUS EVENTS LOG Save Path: C:\Documents and Settings\Administrator\Desktop\ILMN_Info\ImageData

Process Controls

Mismatch

High AT/GC

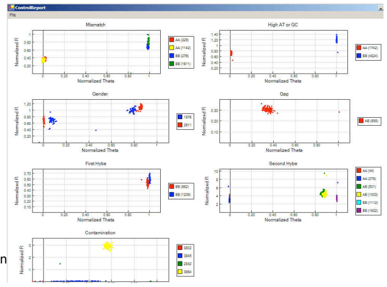
Gender

Gap

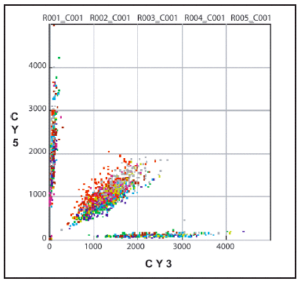
First Hyb

Second Hyb

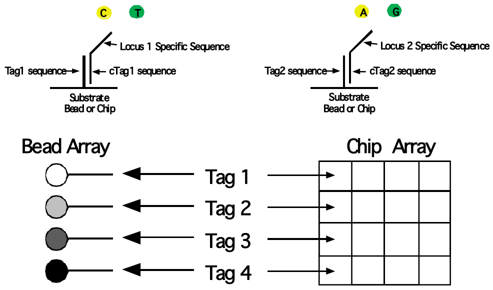
Contamination



Illumina Readout for Sentrix Array
> 1,000 SNPs Assayed on 96 Samples



Multiplexed Genotyping - Universal Tag Readouts



Illumina **Multiplex ~96 - 20,000 SNPs** **ParAllele**

Not dependent on primary PCR **Affymetrix**

Parallele - Defined and Custom Formats

- Intermediate Strategy

- Multiplex ~ 20,000 SNPs

- Affymetrix readout Universal Arrays

Parallele Technology (MIP)

MIP Genotyping Process Overview

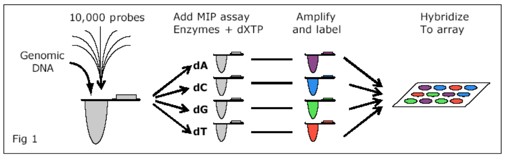
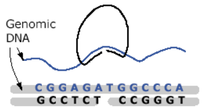


Fig 1

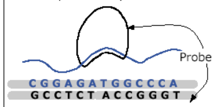
Molecular Inversion Probes (MIP)

1. Anneal



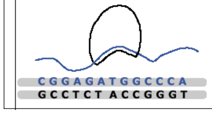
Anneal A mixture of Genomic DNA, up to 10,000 probes, thermostable ligase and polymerase is heat denatured and brought to annealing temperature. Two sequences located at each termini of the probe hybridize to their respective complementary sites on the genome thus forming a circular conformation with a single nucleotide gap between the termini of the probe.

2. Gap Fill - Polymerization



Gap Fill polymerization Unlabeled dATP, dCTP dGTP or dTTP is added to each of the 4 reactions respectively. In reactions where the added nucleotide is complementary to the base being studied, DNA polymerase adds the nucleotide

3. Gap Fill - Ligation



Gap Fill ligation DNA ligase closes the gap to form a covalently closed circular molecule that encircles the genomic strand to which it is hybridized.

4. Exonuclease selection

Exonuclease selection Exonucleases are then added to digest linear probes in reactions where the added nucleotide was not complementary to the gap and excess linear probe in reactions where circular molecules were formed. The reactions are then heated to inactivate the exonucleases.

5. Probe release

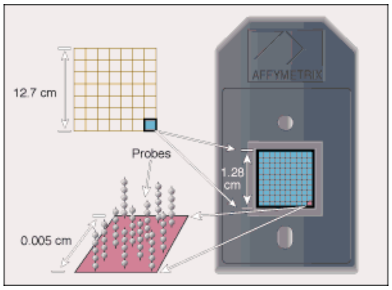
Probe release The probes are then cleaved to release them from the genomic DNA

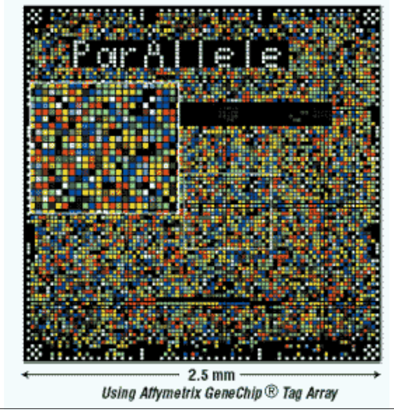
6. Amplification

Amplification The probes are amplified using common primers for all probes

Fig 3

Affymetrix's Chip





Whole Genome Association Strategies

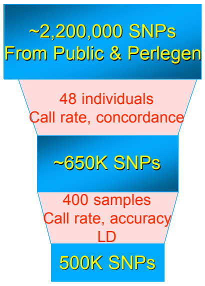
Two Platforms Available Different Designs

- Affymetrix

- Illumina

Affymetrix GeneChip Mapping 500K Array Set

500K: Content Optimized SNP Selection


~2,200,000 SNPs From Public & Perlegen

48 individuals Call rate, concordance

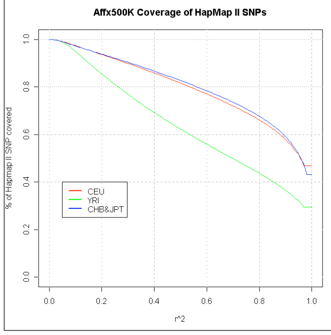
~650K SNPs

400 samples Call rate, accuracy LD

500K SNPs

- Initial Selection: 48 people
  - 2.2M SNPs
  - 25 million genotypes
  - 16 each Caucasian, African, Asian
  - All HapMap samples
- Maximize performance: Second selection over 400 people
  - 270 HapMap Samples
  - 130 diversity samples
  - Accuracy
    - HW, Mendel error, reproducibility
  - Call rates
- Maximize information content:
  - Prioritize SNPs based on LD & HapMap (Broad Institute)

80% genome coverage of Mapping 500K



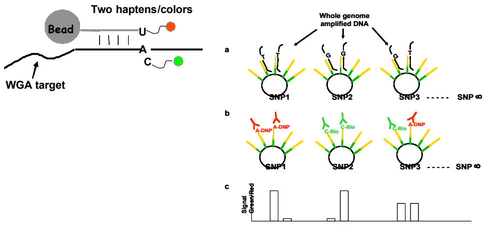
- 500K run on 270 HapMap samples
- Pairwise r^2 analysis for common SNPs (MAF>0.05)
- Robust coverage across populations r^2=0.8
  - CEPH, Asian ~66%
  - Yoruba ~45%
- 2 & 3 marker predictors (multimarker) further increase coverage

Mapping 500K Set

- >500K SNP's
  - 2 array set
- Performance
  - 93-98% call rate range (>95% average)
  - >99.5% concordance with HapMap Genotypes, 99.9% reproducibility
- SNP lists, annotation and genotype data available without restriction at Affymetrix.com

Illumina - Infinium I & II 10K - 300K

Infinium II Assay Single Base Extension

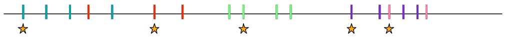


HumanHap-1 Genotyping BeadChip Content

Maximize coverage of human variation by choosing tag SNPs to uniquely identify haplotypes.

Tag SNP selection process:

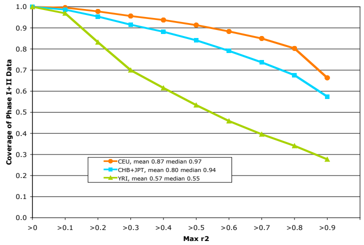
1. Examine HapMap Phase I SNPs with MAF ≥ 0.05 in CEU
2. Bin SNPs in high LD with one another using ldSelect (Carlson, et al. 2004)
3. Select tag SNP with highest design score for each bin.



HumanHap300 Content Strategy

- Tag SNPs
  - $r^2 \geq 0.80$ for bins containing SNPs within 10kb of genes or in evolutionarily conserved regions (ECRs)
  - $r^2 \geq 0.70$ for bins containing SNPs outside of genes or ECRs.
- Additional Content
  - ~8,000 nsSNPs
  - ~1,500 tag SNPs selected from high density SNP data in the MHC region
- Total 317,503 loci

HumanHap300 Genomic Coverage by Population



HumanHap300 Data Quality

127 samples

25 trios

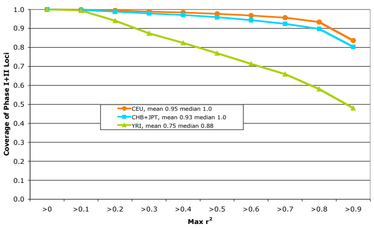
15 replicates

Parameter	Percent
Call rate	99.93%
Reproducibility	>99.99%
Mendelian Inconsistencies	0.035%
Concordance with HapMap Data	99.69%

HumanHap500 Content Strategy

- Analysis of full HapMap data set (Phase I + II) using HumanHap300 SNP list
- Fill in regions of low LD requiring higher density of tag SNPs
- Content Strategy
  - **$r^2 \geq 0.80$ for bins containing SNPs within 10kb of genes or in evolutionarily conserved regions (ECRs) in CEU**
  - **$r^2 \geq 0.70$ for bins containing SNPs outside of genes or ECRs in CEU**
  - **$r^2 \geq 0.80$ for large bins (≥ 3 SNPs) in CHB+JPT population**
  - **$r^2 \geq 0.70$ for large bins (≥ 5 SNPs) in YRI population**

Preliminary HumanHap500 Genomic Coverage by Population



Data Quality Control

- Estimating Error Rates
- Hardy Weinberg Equilibrium
- Frequency Analysis
- Missing Data

Measuring Error Rates

- Genotype replicate samples
- Error rates generally < <1%
- Error rates are SNP specific

		Rep 1		
		CC	CT	TT
Rep 2	CC	24	1	0
	CT	0	50	0
	TT	0	0	25

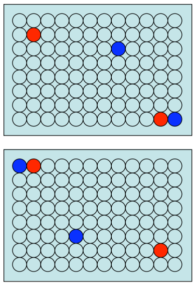
Measuring Error Rates

- Genotype replicate samples
- Absolute number of replicates is more important than percentage
  - E.g. 10% of 200?

		Rep 1		
		CC	CT	TT
Rep 2	CC	24	1	0
	CT	0	50	0
	TT	0	0	25

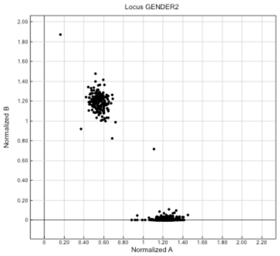
Replicate samples

- Replicates can also detect sample handling errors
  - Wrong plate
  - Plate rotation



Sample Handling Errors

- Sexing samples
- Other known genotypes
  - Blood type
  - HLA
  - Etc.



Hardy Weinberg Equilibrium

- Given
  - p = Allele 1 frequency
  - q = 1-p
- Expectations
  - p^2 = frequency 11
  - 2pq = frequency 12
  - q^2 = frequency 22

Hardy Weinberg Disequilibrium

- Heterozygote excess
  - Biologic
    - Differential survival
  - Technical
    - Nonspecific assays
      Duplicated regions
- Homozygote excess
  - Biologic
    - Population stratification
    - Null allele
  - Technical
    - Allele dropout

HWE Example

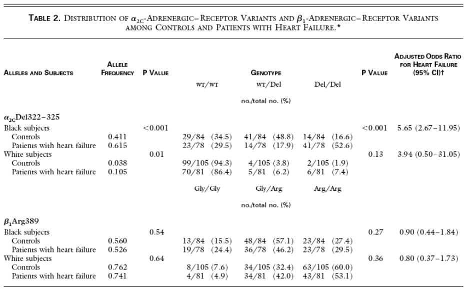
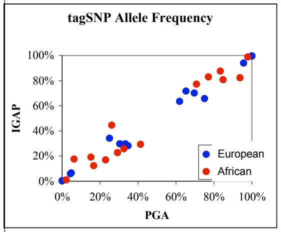


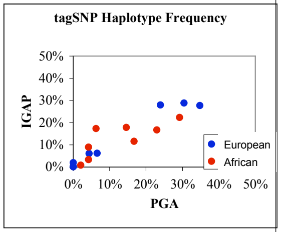
Table 2. Distribution of α_{2C}-Adrenergic–Receptor Variants and β_1-Adrenergic–Receptor Variants among Controls and Patients with Heart Failure.*

Alleles and Subjects	Allele Frequency	P Value	Genotype			P Value	Adjusted Odds Ratio for Heart Failure (95% CI)†
			wt/wt	wt/Del	Del/Del		
			no./total no. (%)				
α_{2C}Del322–325							
Black subjects		<0.001				<0.001	5.65 (2.67–11.95)
Controls	0.411		29/84 (34.5)	41/84 (48.8)	14/84 (16.6)		
Patients with heart failure	0.615		23/78 (29.5)	14/78 (17.9)	41/78 (52.6)		
White subjects		0.01				0.13	3.94 (0.50–31.05)
Controls	0.038		99/105 (94.3)	4/105 (3.8)	2/105 (1.9)		
Patients with heart failure	0.105		70/81 (86.4)	5/81 (6.2)	6/81 (7.4)		
			Gly/Gly	Gly/Arg	Arg/Arg		
			no./total no. (%)				
β_1Arg389							
Black subjects		0.54				0.27	0.90 (0.44–1.84)
Controls	0.560		13/84 (15.5)	48/84 (57.1)	23/84 (27.4)		
Patients with heart failure	0.526		19/78 (24.4)	36/78 (46.2)	23/78 (29.5)		
White subjects		0.64				0.36	0.80 (0.37–1.73)
Controls	0.762		8/105 (7.6)	34/105 (32.4)	63/105 (60.0)		
Patients with heart failure	0.741		4/81 (4.9)	34/81 (42.0)	43/81 (53.1)		

Small et al, NEJM 2002 v347 p1135

Frequency Analysis





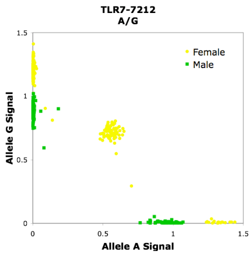
- Check haplotype and allele frequencies against standard populations

Data Quality Control

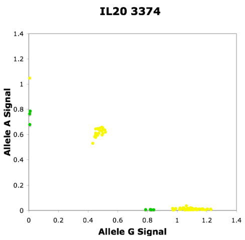
- Estimate Error Rates from Replicates
- Check Hardy Weinberg Equilibrium
- Check Allele and Haplotype Frequencies
- Check Missing Data - Site specific
  Sample specific

Detection of Outliers of the Distribution

X-linked SNP



Unknown SNP



SNP Genotyping Summary

1. Many different genotyping approaches are available - Low to high throughput

2. Some platforms permit users to pick custom SNPs but the highest throughputs are available only in fixed contents.

3. Not all custom SNPs will work for every format. Multiple formats will be required to carry out most projects targeting specific SNPs

4. There are still trade-offs for throughput - Samples vs. SNPs

5. Costs still dictate study design.

6. Regardless of the study - Design,quality control and tracking will rule the day!
Laboratory Information Management Systems are key in every study design
(Key: Track - Samples,
- Assays
- Completion rate
- Reproducibility/Error Analysis)