SNPing in the Human Genome

- Why SNPs?
- How do we find SNPs?
- How many variants?
- What are their frequencies?
- How are they associated with one another?
- How will we apply SNPs?

Why Single Nucleotide Polymorphisms (SNPs)?

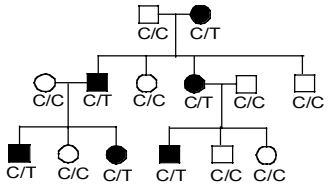
1) Number of Available Markers -

   SNPs 1 in 1,000 bp ~ 3,000,000 Markers

2) Diallelic - Nucleotide Substitutions
3) Number of High Throughput Typing Methods
4) Found in Coding and Non-Coding Regions
5) Potential Association with Functional Variation

Strategies for Genetic Analysis

Linkage Analysis



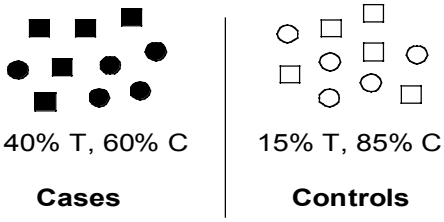
Single Gene

Mendelian Inheritance

Rare (High Penetrance)

~300 Short Tandem Repeat Markers

Association Studies*



Polygenic (also G X E)

Complex Inheritance (Disease susceptibility)

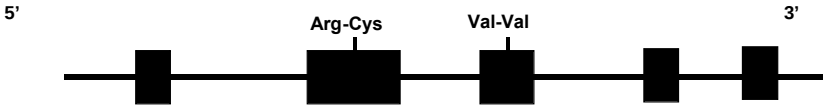
Common

~30,000 to 500,000 or More Polymorphic Markers

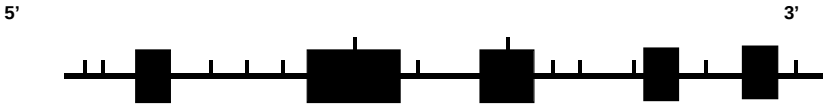
*** Example ApoE4 and Alzheimer's Disease**

Association Studies

Direct:

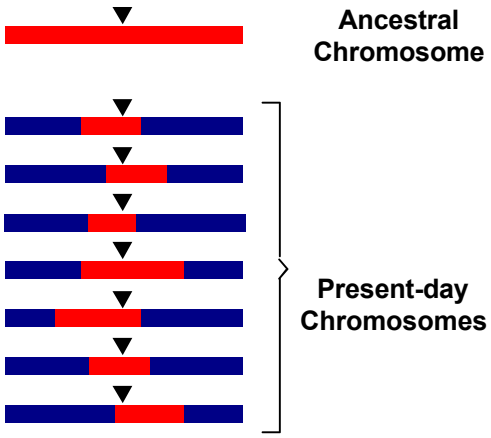


Indirect:



Collins, Guyer, Charkravarti - Science 278: 1580-1581, 1997

Linkage Disequilibrium Around a DNA Variant



From: L. Kruglyak

Identifying DNA Variations

Single-Stranded Conformational Polymorphism (SSCP) Analysis

Enzymatic Cleavage (Restriction Enzymes, RFLPs)

Chemical Cleavage

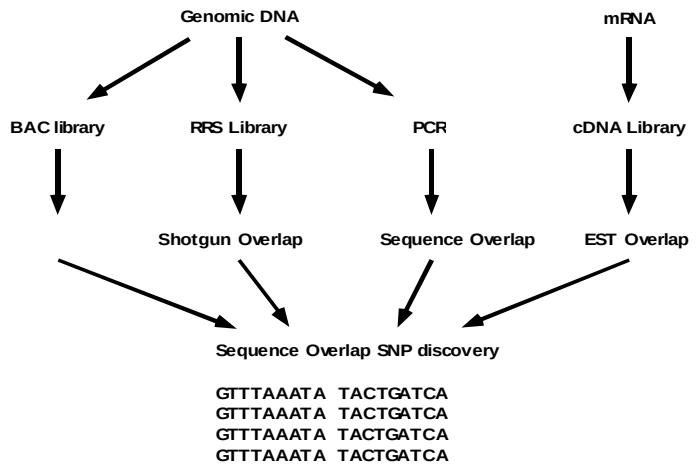
Heteroduplex Analysis

Melting Gradient Analysis (Temperature or Denaturing Chemicals)

Hybridization on Oligonucleotide Microarrays

DNA Sequencing

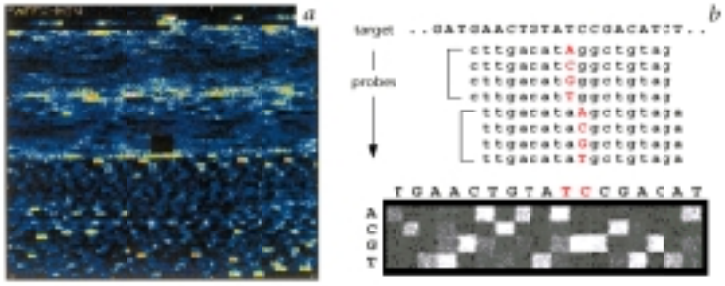
Sequence-based SNP Identification



> 2 million SNPs available: http://www.ncbi.nlm.nih.gov/SNP/

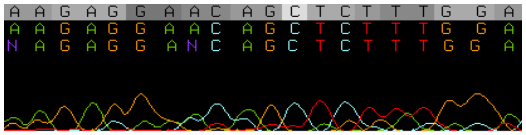
SNP Discovery

Variant Detector Arrays - Sequencing Chips:



Produce High Quality Sequences

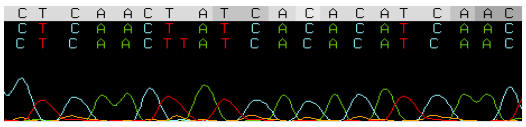
Q = 20



Error estimate:

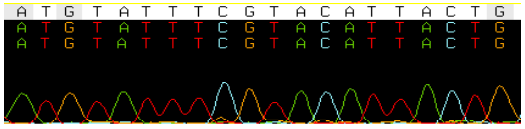
1 in 100

Q = 30



1 in 1,000

Q = 40

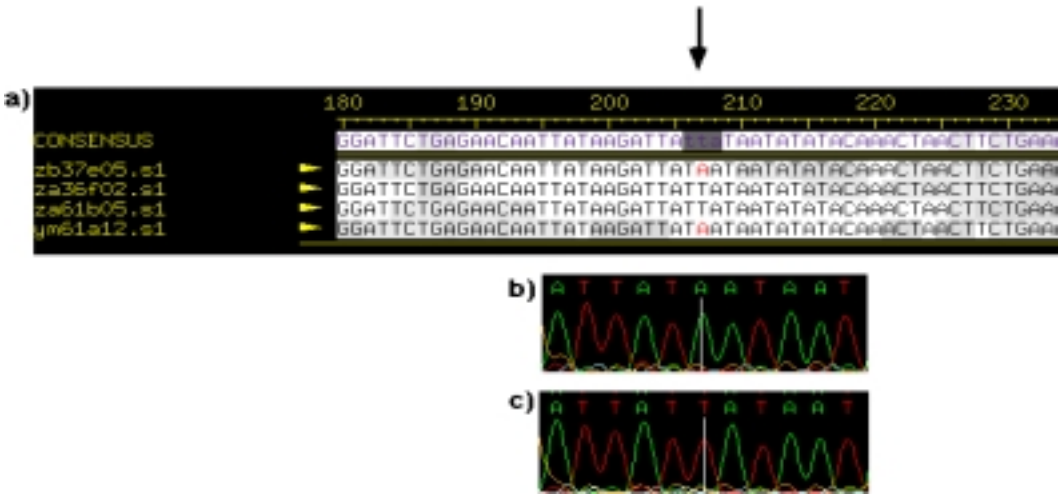


1 in 10,000

$$Q = -10 \log_{10}(p)$$

Phred Quality linked to error estimates
High Quality = High Accuracy

Sequence-based SNP Mining



Detection Rate for SNPs with a Given Minimal Allele Frequency

n	1%	5%	10%	20%	30%	40%
2	.21	.30	.36	.43	.47	.49
3	.32	.46	.55	.65	.71	.74
4	.39	.56	.66	.77	.83	.86
5	.44	.62	.73	.84	.90	.93
6	.48	.68	.78	.89	.94	**.96**
7	.52	.72	.83	.92	**.96**	.98
8	.55	.75	.86	.94	.98	**.99**
9	.57	.78	.88	**.96**	.98	.99
10	.59	.80	.90	.97	**.99**	.997
16	.69	.89	**.96**	**.99**	.999	>.999
24	.76	**.95**	**.99**	.999	>.999	>.999
48	.87	**.99**	.999	>.999	>.999	>.999
96	**.95**	.999	>.999	>.999	>.999	>.999
192	**.99**	>.999	>.999	>.999	>.999	>.999

Occurrence of SNPs in the human population and their representation in the current collection

minimal allele frequency	expected SNP number (millions)	expected SNP frequency (bp)	expected % discovered
1%	11.0	290	11-12
5%	7.1	450	15-17
10%	5.3	600	18-20
20%	3.3	960	21-25
30%	2.0	1570	23-27
40%	0.97	3280	24-28

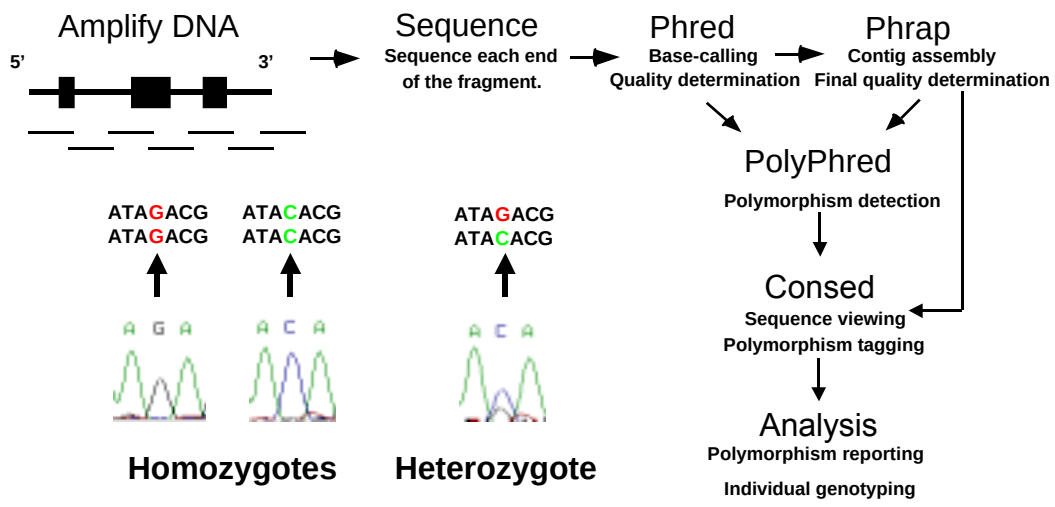
Kruglyak and Nickerson Nature Genetics 27: 234-236

Automating the Analysis of DNA Variations

Issues:

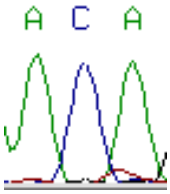
- **Pick primers to amplify and resequence over contiguous regions (10 - 50 kb +)**
- **Produce high quality sequence (i.e. high signal-to-noise)**
- **Automate the detection of DNA variants**
- **Simplify the analysis of the variants**

Sequence-based SNP Identification



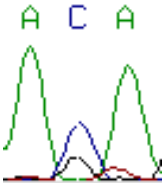
Finding Heterozygotes

Homozygote
2 copies C



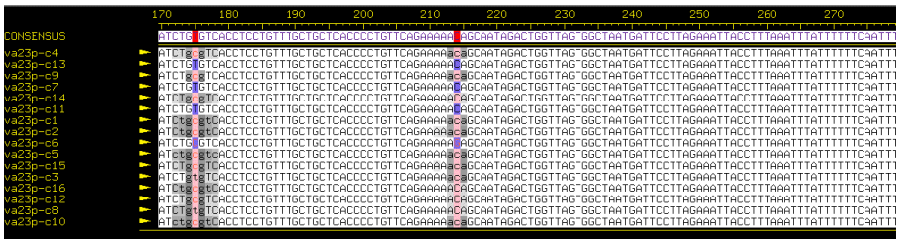
2 X

Heterozygote
1 copy C
1 copy G



1 X

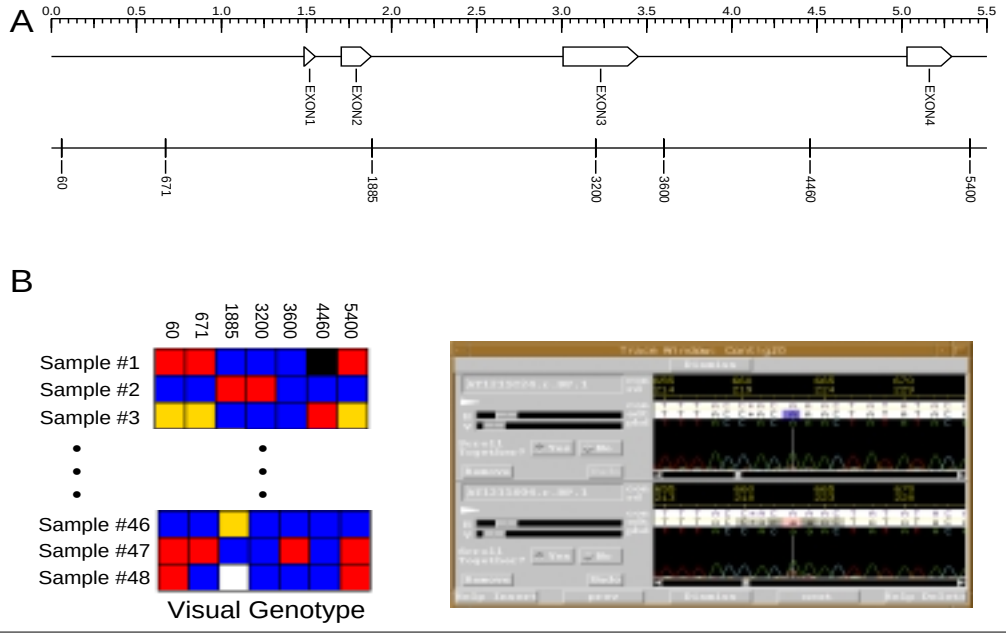
PolyPhred 3.5



RANKINGS 1 2 3 4 5 6

Increased specificity - 10-fold

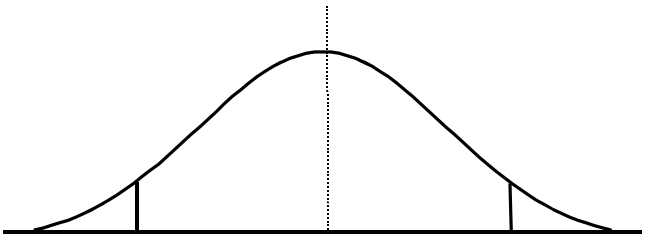
High Throughput SNP Analysis



PolyPhred output provides genotypes for each polymorphic site

```
*****************************************************************
POLYPHRED VERSION        3
*****************************************************************
POLYPHRED_THUMBPRINT     951152661
TIME  21/02/00 09:04:21
BEGIN_COMMAND_LINE
-d                  ../
-ace                vb25new.fasta.screen.ace.1
-quality            20
-ratio              0.65
-background         0.25
-groups             .+
-tag                genotype
-rank               RANK_3
-filter             4
-applyTags          yes
END_COMMAND_LINE
BEGIN_CONTIG
Contig1
BEGIN_SAMPLE
VB25.06.r                1      692    37
VB25.12.r                1      717    35
VB25.11.r                1      591    39
VB25.08.r                1      716    37
VB25.02.r                1      726    35
VB25.04.r                1      717    34
VB25.09.r                1      612    39
VB25.05.r                1      724    38
VB25.10.r                1      596    34
VB25.03.r                1      607    34
VB25.01.r                1      712    37
VB25.07.r                1      690    36
END_SAMPLE
BEGIN_POLY
497       TATTTAACA            C      T      TGTGGATTG          1
699       CCAAAAAAC            C      A      TAACTGTGT          1
877       TTAGAGCAG            A      G      GAGCTGACA          1
END_POLY
BEGIN_GENOTYPE
497       169       VB25.06.r       C    T    1
497       166       VB25.12.r       C    C    1
497       166       VB25.11.r       C    T    1
497       166       VB25.08.r       C    T    1
497       166       VB25.02.r       C    C    1
```

Mean Arterial Pressure(MAP)



Hypertension

- Complex Genetic Trait
- Environmental Factors
- Risk Factor

  CHD, Stroke, Hypertrophy, Atherosclerosis
- Treatment

  Lifestyle - Exercise, Weight reduction, Sodium intake
  Pharmacological - Diuretics, Calcium channel blockers, ACE inhibitors, Angiotensin receptor blockers

Renin-Angiotensin System

Angiotensinogen (15kb)

↓ ← **Renin (8kb)**

Angiotensin I

↓ ← **Angiotensin Converting Enzyme (ACE) (25 kb)**

Angiotensin II

↓ → ↘

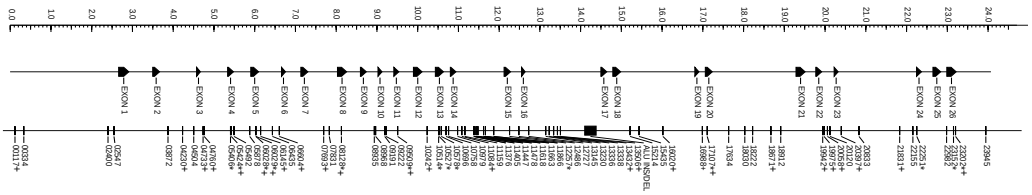
Type-1 receptor (55kb)

Type-2 receptor (5kb)

Mark J. Rieder - UW

Andy Clark - Penn State
Richard Cooper - Loyola

Sequence Diversity in ACE



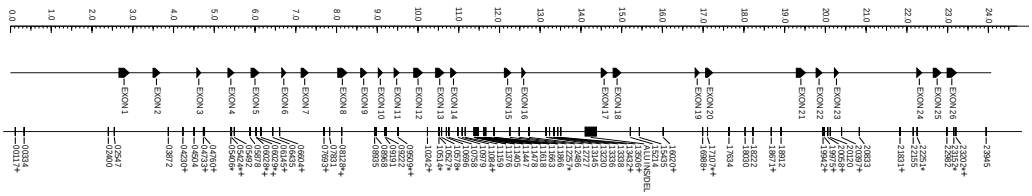
- 78 variant sites - 1 site/1075 bp
  - 63 noncoding
  - 15 coding (5 amino acid changes)

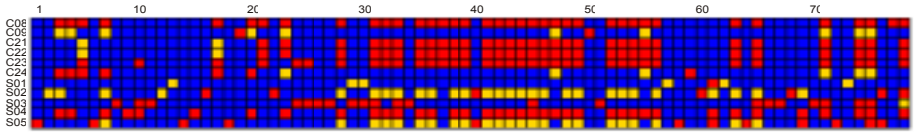
- European-American: 44 varying sites, 4 singletons
- African-American: 70 varying sites, 22 singletons

- No difference in nucleotide diversity between different gene regions or groups within ACE

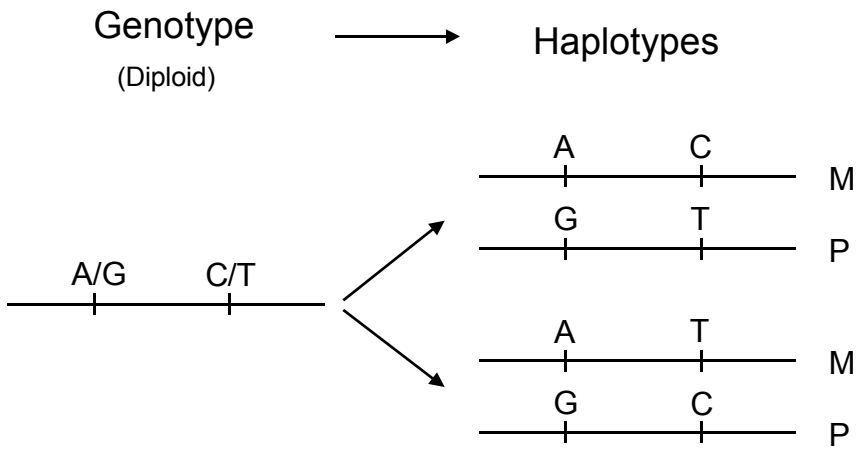
- 18 sites in absolute linkage disequilibrium

Sequence Diversity in ACE

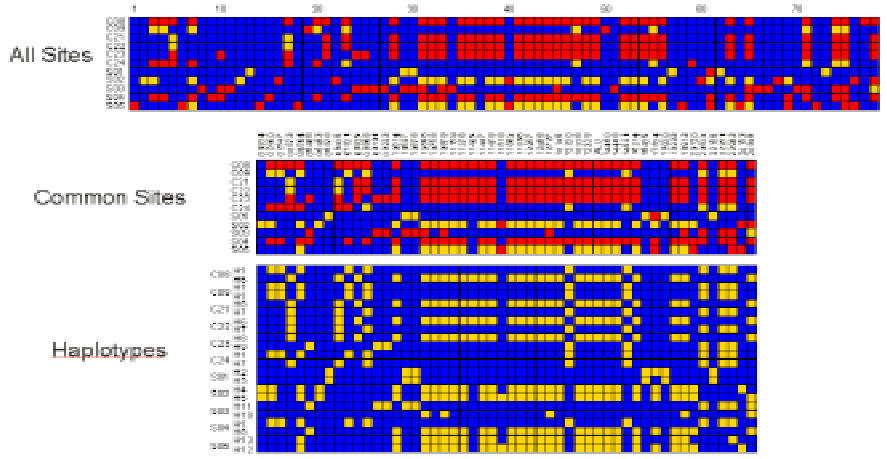




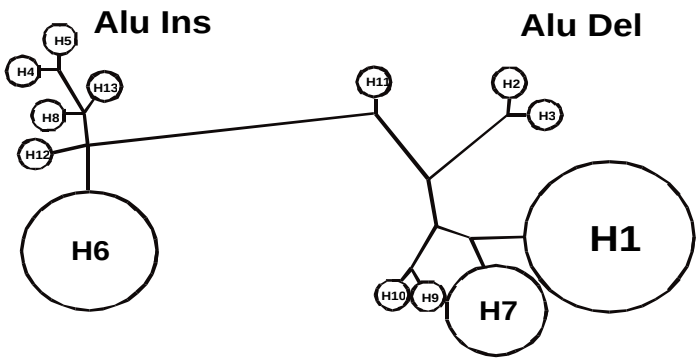
HAPLOTYPING



ACE - Visual Genotypes and Haplotypes



ACE Haplotype Tree



'D' allele - ↑ enzyme activity, hypertension, LVH, CAD, renal disease, longevity
'I' allele - ↑ LV mass, elite endurance/athletic performance

Serum ACE Activity and ACE Genotype

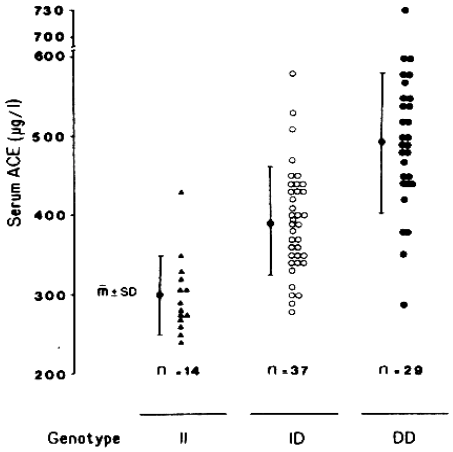
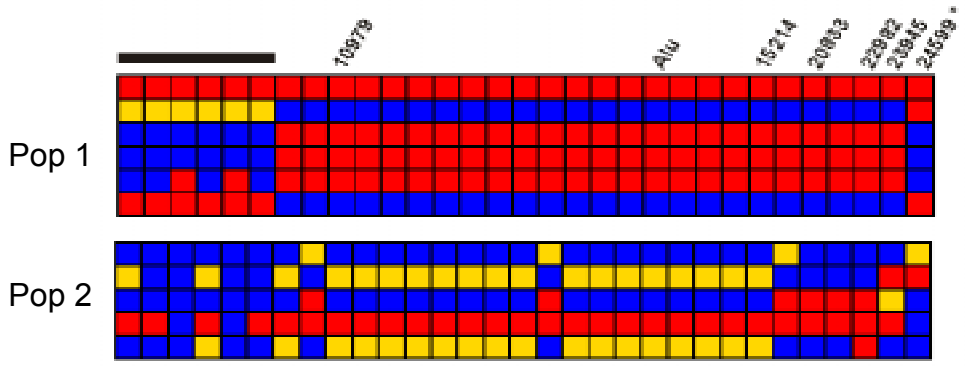


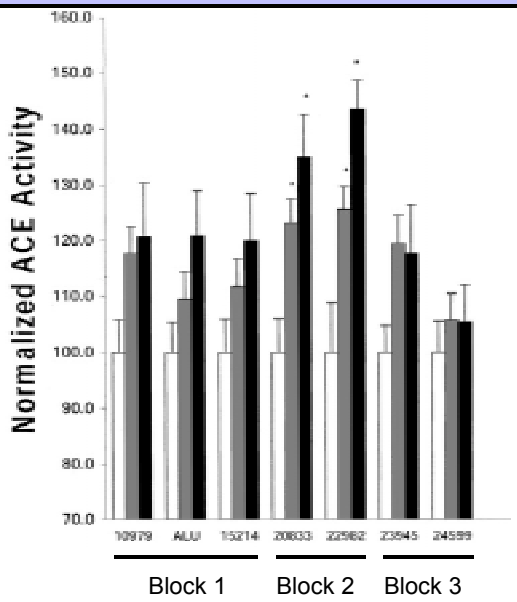
Figure 2. Serum immunoreactive ACE concentrations (μg/liter) for individual with the II, ID, and DD genotypes, respectively, shown in left, middle, and right panels. Solid vertical bars indicate mean concentration and standard deviation for each group.

Rigat et al. J. Clin. Invest. 1990

Common ACE SNPs

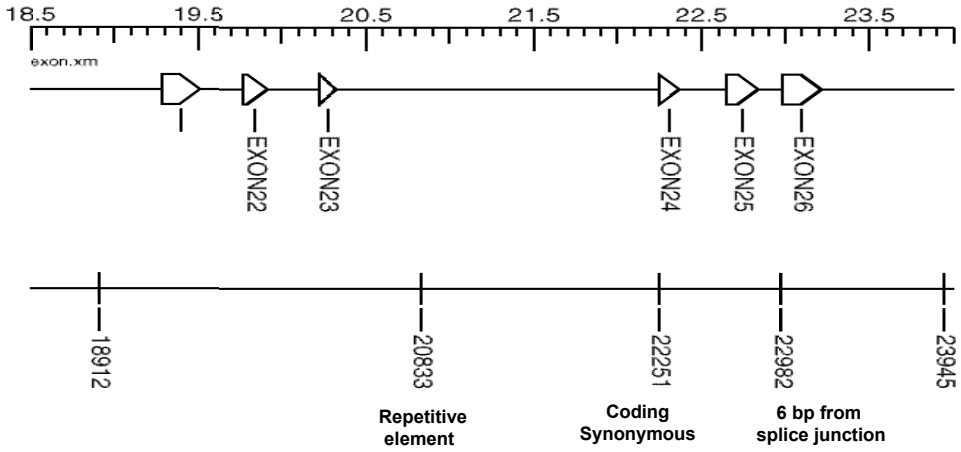


Fine-Mapping ACE Activity



Zhu et al
Am. J. Hum. Genet.
67: 1144-1153, 2000

Functional Variants



Possibilites: **Alternative splicing - Sugimura et al. BBRC 247:466, 1998**
Synonymous variation - Translational efficiency - Shen et al. PNAS 96: 7871, 1999
Non-coding variation - Regulatory elements- Loots et al. Science 288: 136, 2000

SNP Mapping

- Many Common Variant Sites
- Organization and History of the Variation is Complex
- Site Selection is an Important Issue