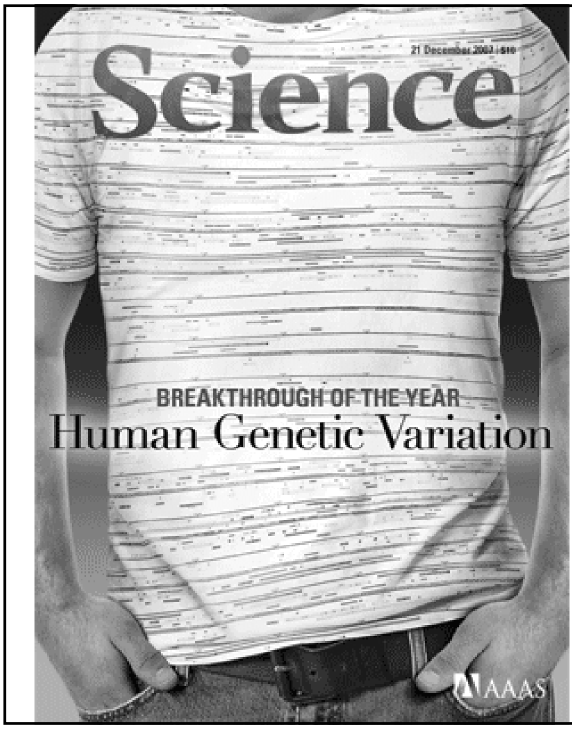
Studying Genetic Variation II: Computational Techniques

Jim Mullikin, PhD

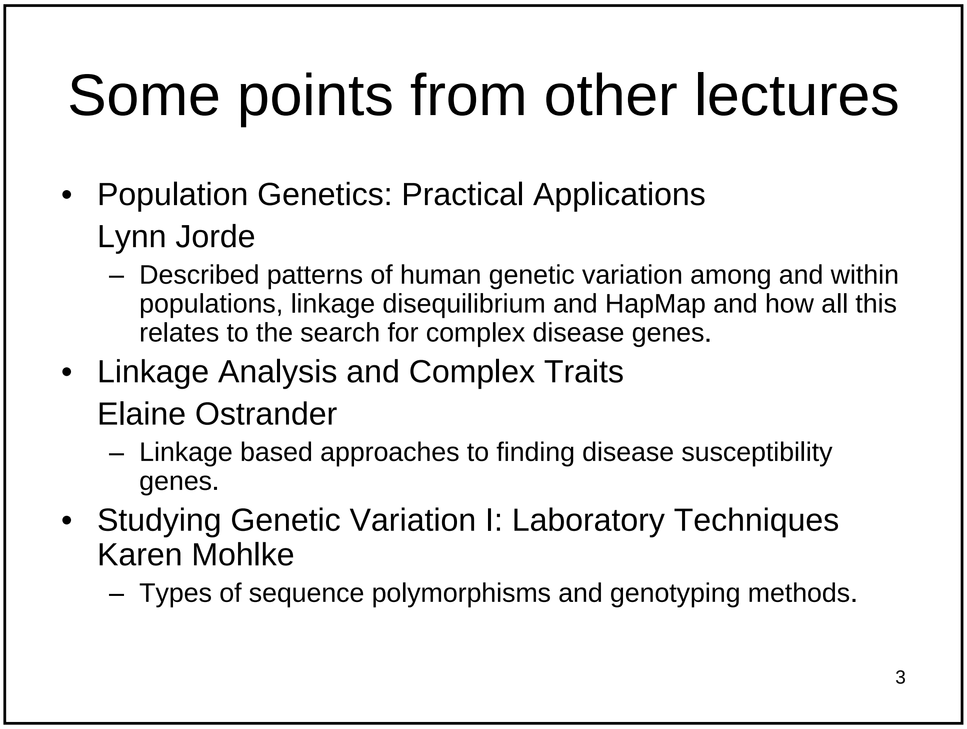
Genome Technology Branch

NHGRI



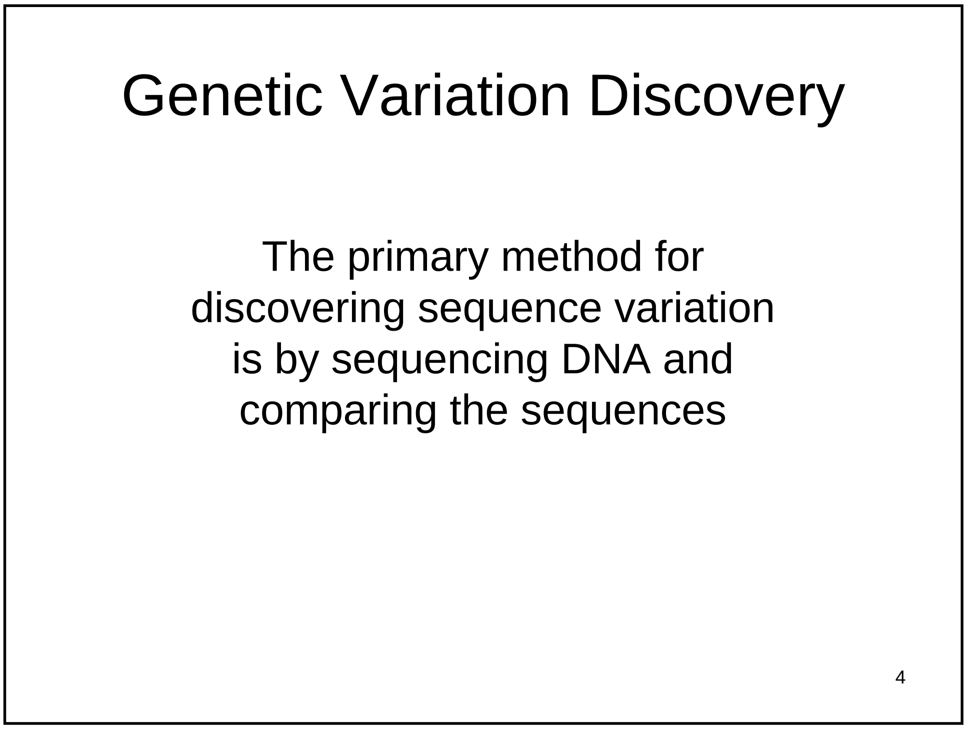
Science

December 21, 2007



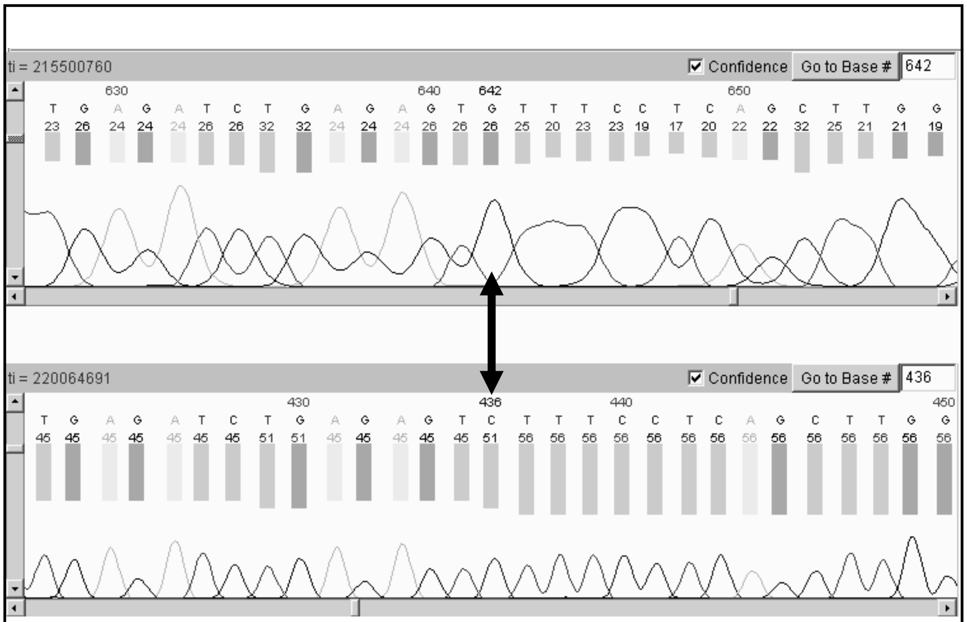
Some points from other lectures

- Population Genetics: Practical Applications
  Lynn Jorde
  - Described patterns of human genetic variation among and within populations, linkage disequilibrium and HapMap and how all this relates to the search for complex disease genes.
- Linkage Analysis and Complex Traits
  Elaine Ostrander
  - Linkage based approaches to finding disease susceptibility genes.
- Studying Genetic Variation I: Laboratory Techniques
  Karen Mohlke
  - Types of sequence polymorphisms and genotyping methods.



Genetic Variation Discovery

The primary method for
discovering sequence variation
is by sequencing DNA and
comparing the sequences



Trace Archive Search query:
trace_name='S213P602260RC9.T0' OR trace_name='50766946'

Overview of Topics

- Review of genetic variation discovery
- Database of SNPs, dbSNP
- Other types of genetic variation
- Medical sequencing
- Next-generation sequencing and SNPs
- Targeted Genomic Selection

A few definitions

- Alleles
  - Alternate forms of a gene or chromosomal locus that differ in DNA sequence
- Single Nucleotide Polymorphism (SNP)
  - The most common form of genetic variation in the genome: a single-base substitution
- Minor Allele Frequency (MAF)
  - Proportion of the less common of 2 alleles in a population
- Polymorphic
  - Usually implies a MAF of at least 1%

NCBI dbSNP database of genetic variation

- http://www.ncbi.nlm.nih.gov/SNP/

- This is the main repository of publicly available genetic variation data.

- You'll also find information on allele frequencies, populations, genotype assays and much more.

Review of Genetic Variation Discovery Efforts

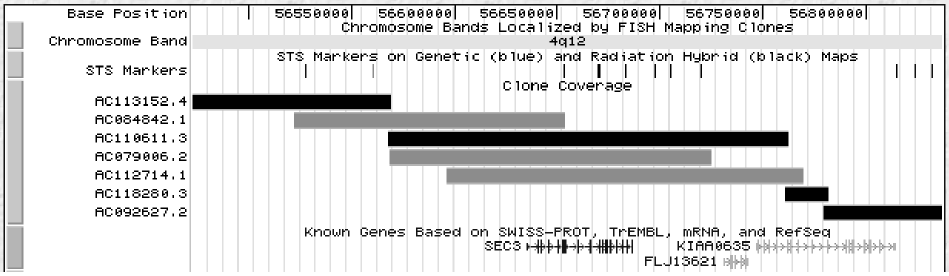
- Expressed sequence tag (EST) mining
- Clone overlap
- The SNP Consortium (TSC)
- Haplotype Map Project (HapMap)
- Chip based sequencing arrays
- Human Genome Structural Variation (HGSV)
- Personal Genomes (available from NCBI trace archive)
  - Craig Venter (*PLoS Biology* Vol. 5, No. 10, e254)
  - Jim Watson (http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/)

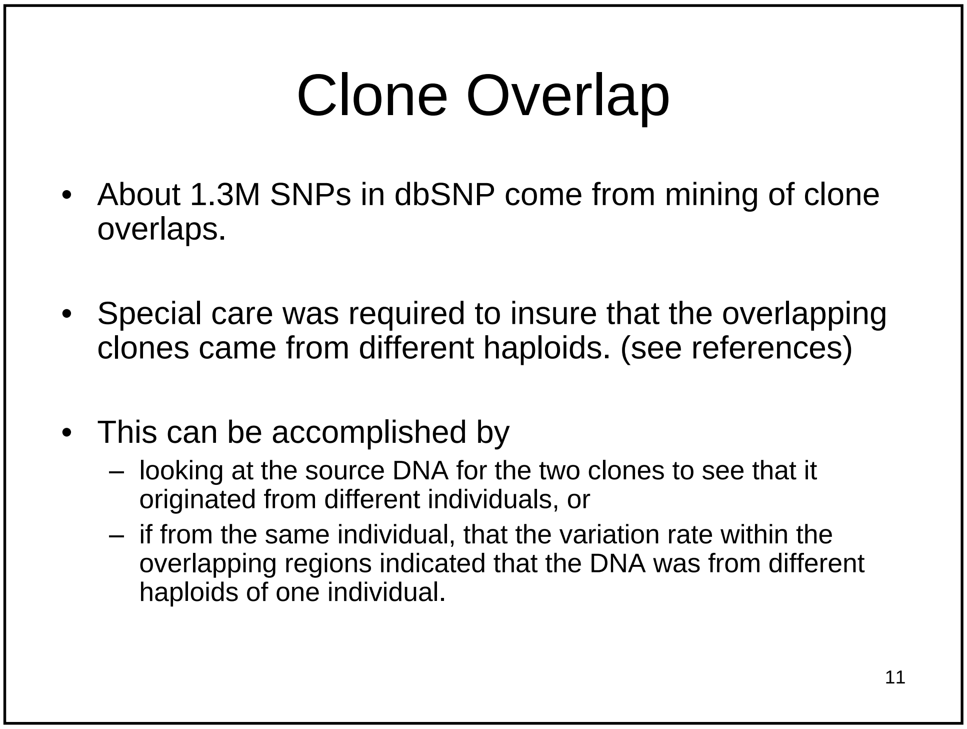
9

Clone Overlap

- The human genome was sequenced from BAC clones (containing about 150kb of sequence each).

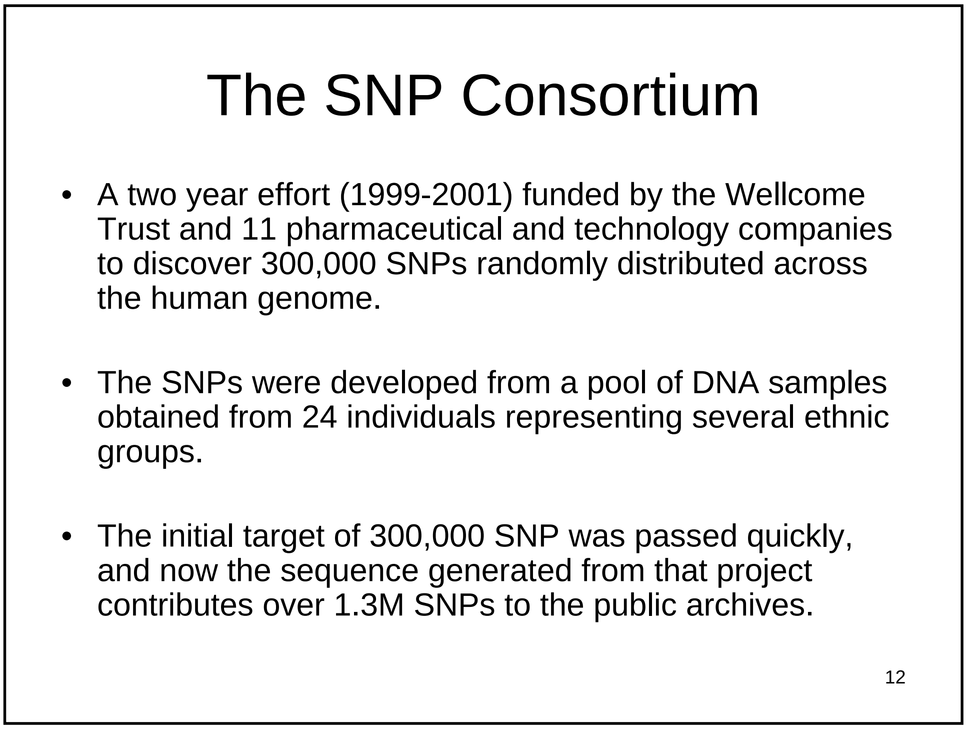
- These overlapped to various levels, and within the overlap regions, high quality base differences indicated the position and alleles of SNPs.





Clone Overlap

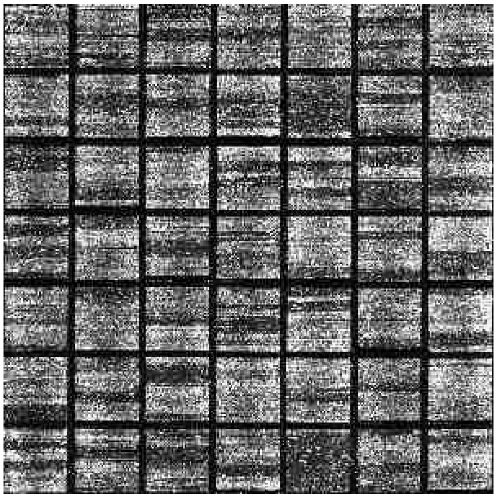
- About 1.3M SNPs in dbSNP come from mining of clone overlaps.
- Special care was required to insure that the overlapping clones came from different haploids. (see references)
- This can be accomplished by
  - looking at the source DNA for the two clones to see that it originated from different individuals, or
  - if from the same individual, that the variation rate within the overlapping regions indicated that the DNA was from different haploids of one individual.



The SNP Consortium

- A two year effort (1999-2001) funded by the Wellcome Trust and 11 pharmaceutical and technology companies to discover 300,000 SNPs randomly distributed across the human genome.
- The SNPs were developed from a pool of DNA samples obtained from 24 individuals representing several ethnic groups.
- The initial target of 300,000 SNP was passed quickly, and now the sequence generated from that project contributes over 1.3M SNPs to the public archives.

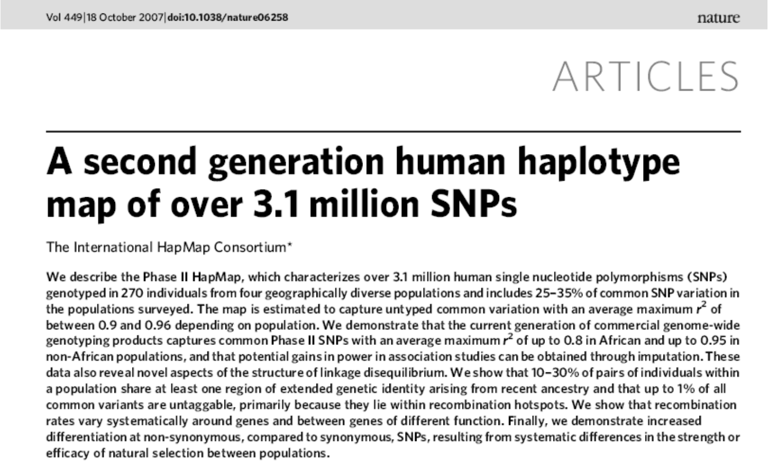
Perlegen used Affymetrix’s chip design process to place 60M probes on a 5x5” chip. From 20 single haploid chromosome 21 chromosomes, they discovered 36k SNPs.



http://www.perlegen.com/

More SNPs for HapMap Project

- This project required many more SNPs than were available when it started in October 2002, which totaled about 2M.
- Additional random shotgun sequencing has brought this to 8.2M SNPs for the HapMap Project.
- It has been estimated that there are perhaps 10M common SNPs (> 5% MAF), so there are many more SNPs yet to discover.



Vol 449|18 October 2007|**doi:10.1038/nature06258** nature

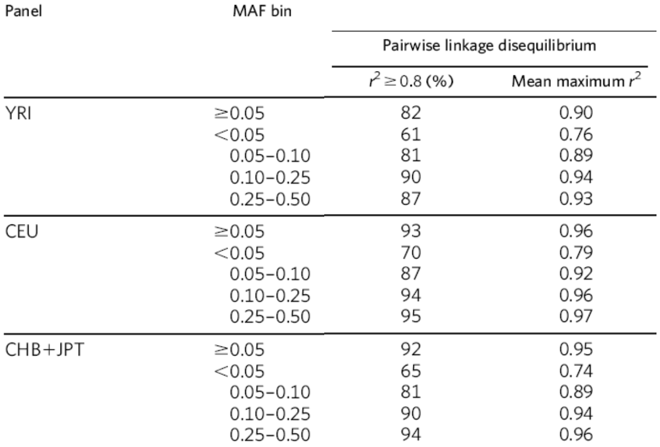
ARTICLES

A second generation human haplotype map of over 3.1 million SNPs

The International HapMap Consortium*

We describe the Phase II HapMap, which characterizes over 3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from four geographically diverse populations and includes 25–35% of common SNP variation in the populations surveyed. The map is estimated to capture untyped common variation with an average maximum r^2 of between 0.9 and 0.96 depending on population. We demonstrate that the current generation of commercial genome-wide genotyping products captures common Phase II SNPs with an average maximum r^2 of up to 0.8 in African and up to 0.95 in non-African populations, and that potential gains in power in association studies can be obtained through imputation. These data also reveal novel aspects of the structure of linkage disequilibrium. We show that 10–30% of pairs of individuals within a population share at least one region of extended genetic identity arising from recent ancestry and that up to 1% of all common variants are untaggable, primarily because they lie within recombination hotspots. We show that recombination rates vary systematically around genes and between genes of different function. Finally, we demonstrate increased differentiation at non-synonymous, compared to synonymous, SNPs, resulting from systematic differences in the strength or efficacy of natural selection between populations.

Table 2. Estimated coverage of the Phase II HapMap in the ten HapMap ENCODE regions



Panel	MAF bin	Pairwise linkage disequilibrium	
		$r^2 \geq 0.8$ (%)	Mean maximum r^2
YRI	≥0.05	82	0.90
	<0.05	61	0.76
	0.05–0.10	81	0.89
	0.10–0.25	90	0.94
	0.25–0.50	87	0.93
CEU	≥0.05	93	0.96
	<0.05	70	0.79
	0.05–0.10	87	0.92
	0.10–0.25	94	0.96
	0.25–0.50	95	0.97
CHB+JPT	≥0.05	92	0.95
	<0.05	65	0.74
	0.05–0.10	81	0.89
	0.10–0.25	90	0.94
	0.25–0.50	94	0.96

NATURE Vol 449, 18 October 2007

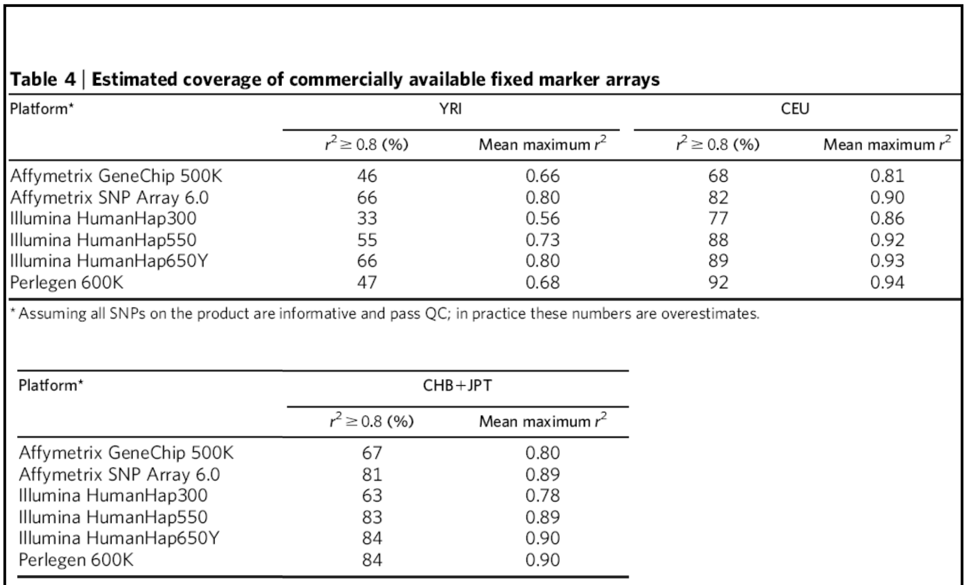


Table 4 | Estimated coverage of commercially available fixed marker arrays

Platform*	YRI		CEU	
	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2
Affymetrix GeneChip 500K	46	0.66	68	0.81
Affymetrix SNP Array 6.0	66	0.80	82	0.90
Illumina HumanHap300	33	0.56	77	0.86
Illumina HumanHap550	55	0.73	88	0.92
Illumina HumanHap650Y	66	0.80	89	0.93
Perlegen 600K	47	0.68	92	0.94

* Assuming all SNPs on the product are informative and pass QC; in practice these numbers are overestimates.

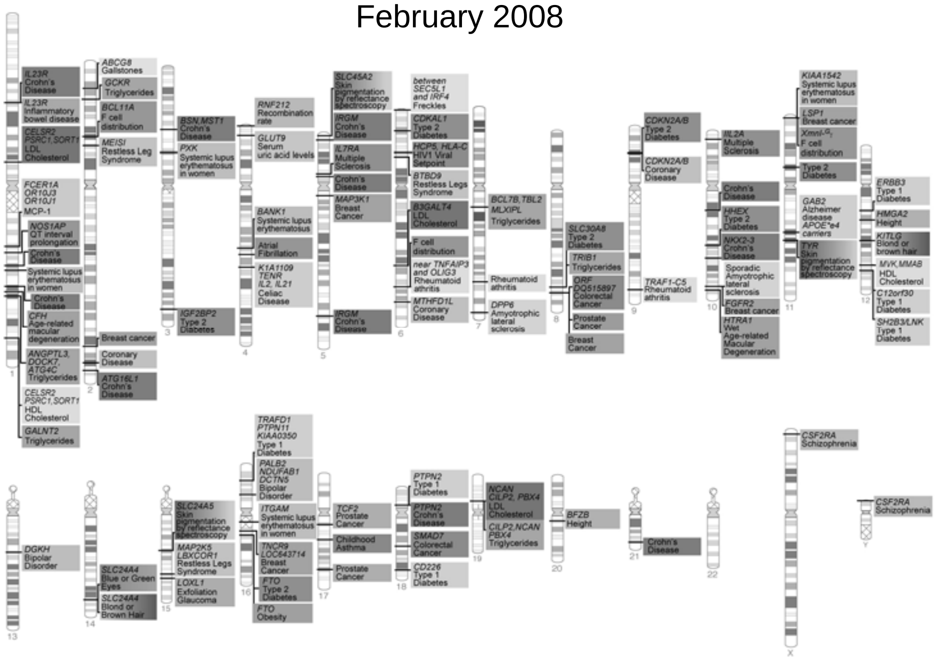
Platform*	CHB+JPT	
	$r^2 \geq 0.8$ (%)	Mean maximum r^2
Affymetrix GeneChip 500K	67	0.80
Affymetrix SNP Array 6.0	81	0.89
Illumina HumanHap300	63	0.78
Illumina HumanHap550	83	0.89
Illumina HumanHap650Y	84	0.90
Perlegen 600K	84	0.90

NATURE Vol 449, 18 October 2007

Genome-Wide Association Studies

- Enabled by the HapMap project and spinoff SNP genotyping chips
- Availability of large, well studied sample cohorts
- Funded internationally
  - Genetic Association Information Network (GAIN, a public-private partnership)
    - http://www.fnih.org/GAIN2/home_new.shtml
  - Genes, Environment and Health Initiative (GEI)
    - http://www.genesandenvironment.nih.gov/
  - Wellcome Trust Case Control Consortium (WTCCC)
    - http://www.wtccc.org.uk/

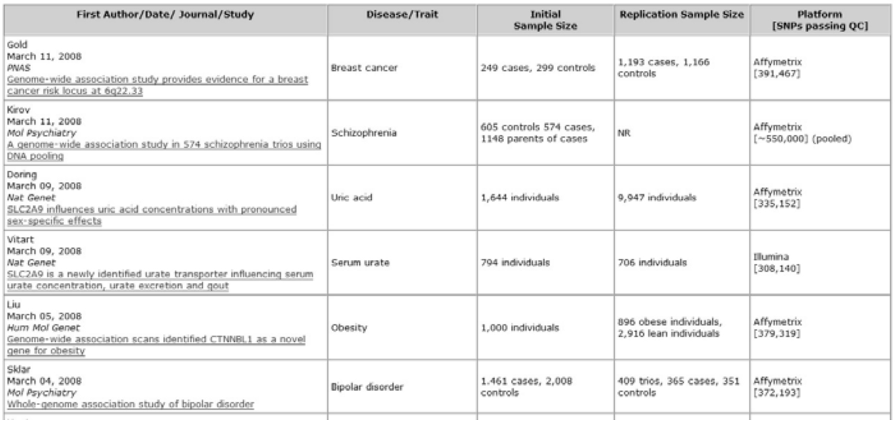
February 2008


IL23R Crohn's Disease
IL23R Inflammatory bowel disease
CELSR2 PSRC1,SORT1 LDL Cholesterol
FCER1A OR10J3 OR10J1 MCP-1
NOS1AP QT interval prolongation
Crohn's Disease
Systemic lupus erythematosus in women
Crohn's Disease
CFH Age-related macular degeneration
ANGPTL3, DOCK7, ATG4C Triglycerides
CELSR2 PSRC1,SORT1 HDL Cholesterol
GALNT2 Triglycerides
1
ABCG8 Gallstones
GCKR Triglycerides
BCL11A F cell distribution
MEISI Restless Leg Syndrome
Breast cancer
Coronary Disease
ATG16L1 Crohn's Disease
2
BSN,MST1 Crohn's Disease
PXK Systemic lupus erythematosus in women
IGF2BP2 Type 2 Diabetes
3
RNF212 Recombination rate
GLUT9 Serum uric acid levels
BANK1 Systemic lupus erythematosus
Atrial Fibrillation
K1A1109 TENR IL2, IL21 Celiac Disease
4
SLC45A2 Skin pigmentation by reflectance spectroscopy
IRGM Crohn's Disease
IL7RA Multiple Sclerosis
Crohn's Disease
MAP3K1 Breast Cancer
IRGM Crohn's Disease
5
between SEC5L1 and IRF4 Freckles
CDKAL1 Type 2 Diabetes
HCP5, HLA-C HIV1 Viral Setpoint
BTBD9 Restless Legs Syndrome
B3GALT4 LDL Cholesterol
F cell distribution
near TNFAIP3 and OLIG3 Rheumatoid athritis
MTHFD1L Coronary Disease
6
BCL7B,TBL2 MLXIPL Triglycerides
Rheumatoid athritis
DPP6 Amyotrophic lateral sclerosis
7
SLC30A8 Type 2 Diabetes
TRIB1 Triglycerides
ORF DQ515897 Colorectal Cancer
Prostate Cancer
Breast Cancer
8
CDKN2A/B Type 2 Diabetes
CDKN2A/B Coronary Disease
TRAF1-C5 Rheumatoid athritis
9
IIL2A Multiple Sclerosis
Crohn's Disease
HHEX Type 2 Diabetes
NKX2-3 Crohn's Disease
Sporadic Amyotrophic lateral sclerosis
FGFR2 Breast cancer
HTRA1 Wet Age-related Macular Degeneration
10
KIAA1542 Systemic lupus erythematosus in women
LSP1 Breast cancer
Xmnl-G F cell distribution
Type 2 Diabetes
GAB2 Alzheimer disease APOE*e4 carriers
TYR Skin pigmentation by reflectance spectroscopy
11
ERBB3 Type 1 Diabetes
HMGA2 Height
KITLG Blond or brown hair
MVK,MMAB HDL Cholesterol
C12orf30 Type 1 Diabetes
SH2B3/LNK Type 1 Diabetes
12
DGKH Bipolar Disorder
13
SLC24A4 Blue or Green Eyes
SLC24A4 Blond or Brown Hair
14
SLC24A5 Skin pigmentation by reflectance spectroscopy
MAP2K5 LBXCOR1 Restless Legs Syndrome
LOXL1 Exfoliation Glaucoma
15
TRAFD1 PTPN11 KIAA0350 Type 1 Diabetes
PALB2 NDUFAB1 DCTN5 Bipolar Disorder
ITGAM Systemic lupus erythematosus in women
TNCR9 LOC643714 Breast Cancer
FTO Type 2 Diabetes
FTO Obesity
16
TCF2 Prostate Cancer
Childhood Asthma
Prostate Cancer
17
PTPN2 Type 1 Diabetes
PTPN2 Crohn's Disease
SMAD7 Colorectal Cancer
CD226 Type 1 Diabetes
18
NCAN CILP2, PBX4 LDL Cholesterol
CILP2,NCAN PBX4 Triglycerides
19
BFZB Height
20
Crohn's Disease
21
22
CSF2RA Schizophrenia
X
CSF2RA Schizophrenia
Y

Manolio, Brooks, Collins, J. Clin. Invest., in press.

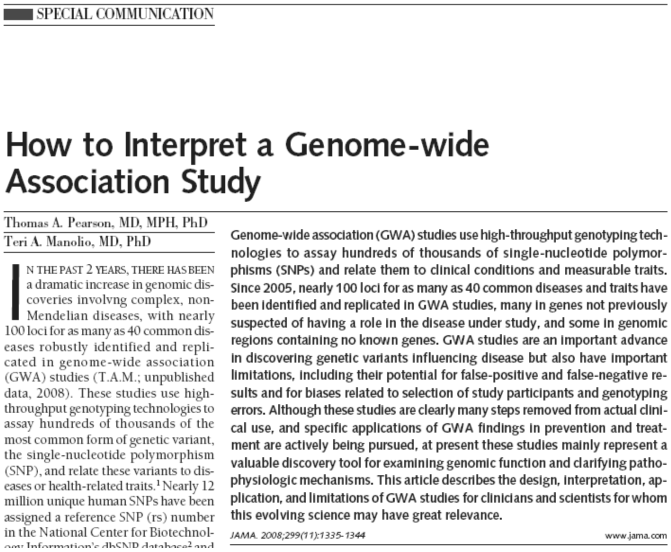
A Catalog of Published Genome-Wide Association Studies

- http://www.genome.gov/26525384



First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Platform [SNPs passing QC]
Gold March 11, 2008 *PNAS* Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33	Breast cancer	249 cases, 299 controls	1,193 cases, 1,166 controls	Affymetrix [391,467]
Kirov March 11, 2008 *Mol Psychiatry* A genome-wide association study in 574 schizophrenia trios using DNA pooling	Schizophrenia	605 controls 574 cases, 1148 parents of cases	NR	Affymetrix [~550,000] (pooled)
Doring March 09, 2008 *Nat Genet* SLC2A9 influences uric acid concentrations with pronounced sex-specific effects	Uric acid	1,644 individuals	9,947 individuals	Affymetrix [335,152]
Vitart March 09, 2008 *Nat Genet* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout	Serum urate	794 individuals	706 individuals	Illumina [308,140]
Liu March 05, 2008 *Hum Mol Genet* Genome-wide association scans identified CTNNBL1 as a novel gene for obesity	Obesity	1,000 individuals	896 obese individuals, 2,916 lean individuals	Affymetrix [379,319]
Sklar March 04, 2008 *Mol Psychiatry* Whole-genome association study of bipolar disorder	Bipolar disorder	1.461 cases, 2,008 controls	409 trios, 365 cases, 351 controls	Affymetrix [372,193]

And 130 more entries…



SPECIAL COMMUNICATION

How to Interpret a Genome-wide Association Study

Thomas A. Pearson, MD, MPH, PhD
Teri A. Manolio, MD, PhD

Genome-wide association (GWA) studies use high-throughput genotyping technologies to assay hundreds of thousands of single-nucleotide polymorphisms (SNPs) and relate them to clinical conditions and measurable traits. Since 2005, nearly 100 loci for as many as 40 common diseases and traits have been identified and replicated in GWA studies, many in genes not previously suspected of having a role in the disease under study, and some in genomic regions containing no known genes. GWA studies are an important advance in discovering genetic variants influencing disease but also have important limitations, including their potential for false-positive and false-negative results and for biases related to selection of study participants and genotyping errors. Although these studies are clearly many steps removed from actual clinical use, and specific applications of GWA findings in prevention and treatment are actively being pursued, at present these studies mainly represent a valuable discovery tool for examining genomic function and clarifying pathophysiologic mechanisms. This article describes the design, interpretation, application, and limitations of GWA studies for clinicians and scientists for whom this evolving science may have great relevance.

JAMA. 2008;299(11):1335-1344 www.jama.com

IN THE PAST 2 YEARS, THERE HAS BEEN a dramatic increase in genomic discoveries involvng complex, non-Mendelian diseases, with nearly 100 loci for as many as 40 common diseases robustly identified and replicated in genome-wide association (GWA) studies (T.A.M.; unpublished data, 2008). These studies use high-throughput genotyping technologies to assay hundreds of thousands of the most common form of genetic variant, the single-nucleotide polymorphism (SNP), and relate these variants to diseases or health-related traits.[1] Nearly 12 million unique human SNPs have been assigned a reference SNP (rs) number in the National Center for Biotechnology Information's dbSNP database[2] and

JAMA. 2008;299(11):1335-1344.

dbGaP

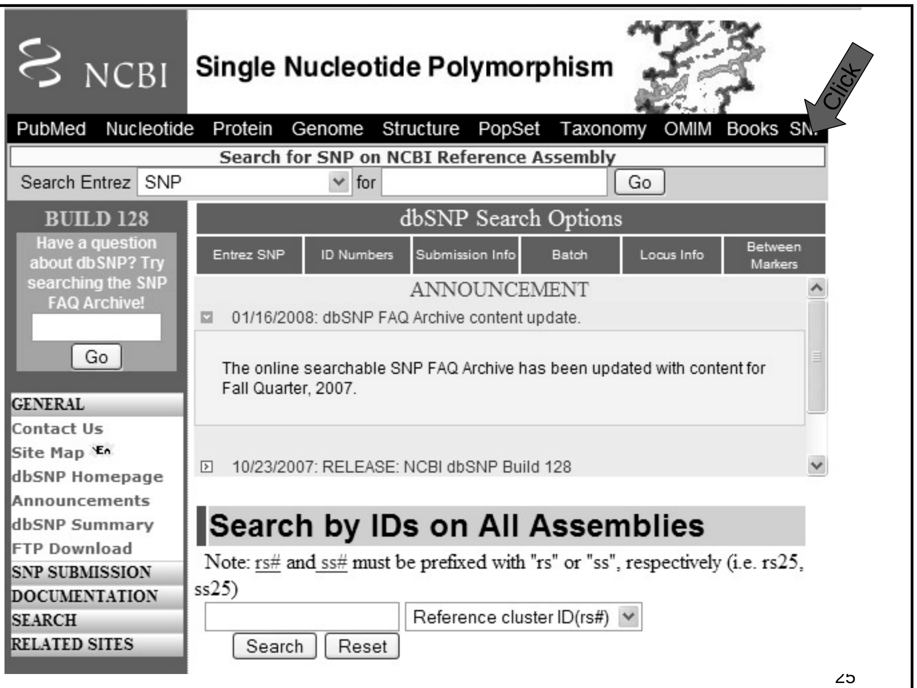
- http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap_tmpl/about.html
- The **d**ata**b**ase of **G**enotype **a**nd **P**henotype (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.
- http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap_tmpl/dbGaP_HowTo.pdf

Overview of Topics

- Review of genetic variation discovery
- Database of SNPs, dbSNP
- Other types of genetic variation
- Medical sequencing
- Next-generation sequencing and SNPs
- Targeted Genomic Selection

What's recorded in dbSNP

- From their main web page, they have extensive information on how to submit SNPs, genotypes, validation experiments, population frequencies, etc., for any species.
- SNPs that you submit are called Submitter SNPs and get ssIDs.
- If there is a reference sequence available for the species submitted, they will map SNPs to this reference using the flank information you provide.
- SNPs that cluster at the same locus, are merged into Reference SNPs which have unique rsIDs.



Search for SNP on NCBI Reference Assembly

Have a question about dbSNP? Try searching the SNP FAQ Archive!

ANNOUNCEMENT

01/16/2008: dbSNP FAQ Archive content update.

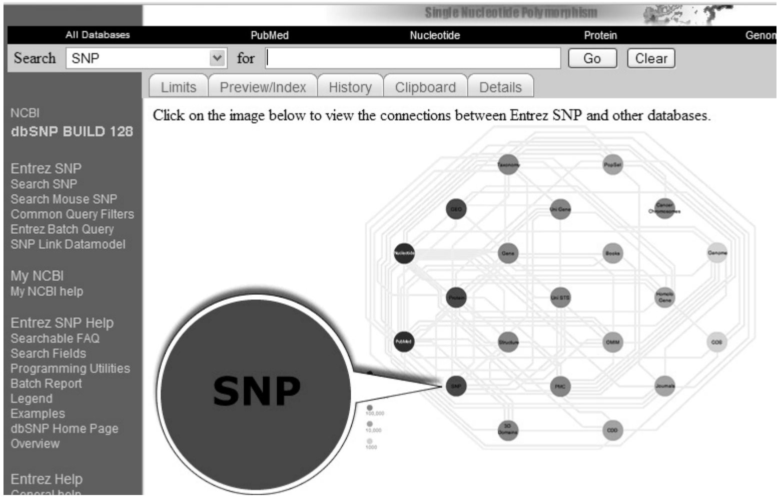
The online searchable SNP FAQ Archive has been updated with content for Fall Quarter, 2007.

10/23/2007: RELEASE: NCBI dbSNP Build 128

Search by IDs on All Assemblies

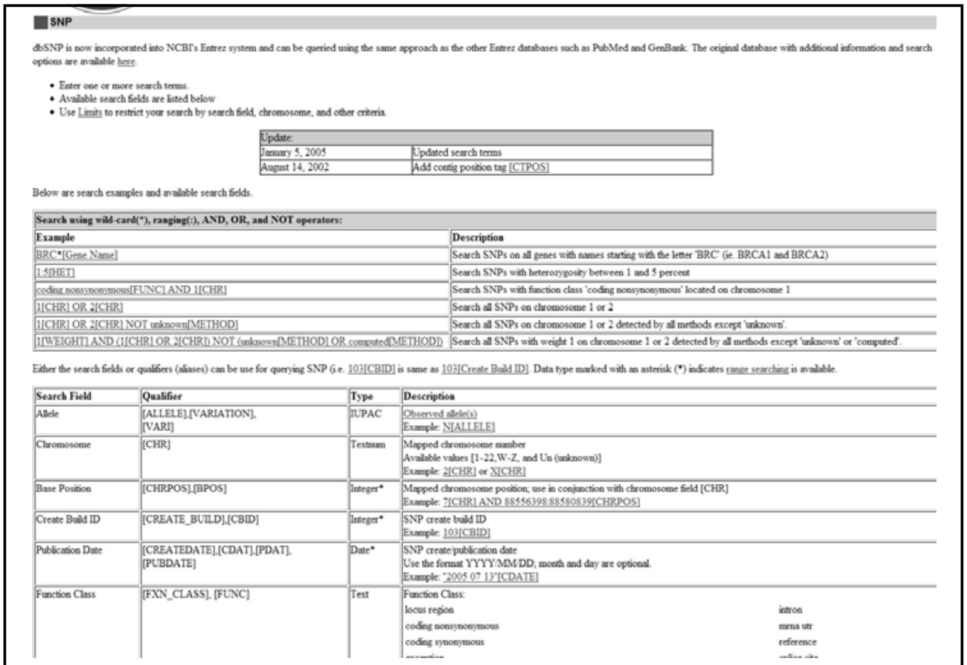
Note: rs# and ss# must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss25)

http://www.ncbi.nlm.nih.gov/SNP/index.html



Click on the image below to view the connections between Entrez SNP and other databases.

http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp



SNP

dbSNP is now incorporated into NCBI's Entrez system and can be queried using the same approach as the other Entrez databases such as PubMed and GenBank. The original database with additional information and search options are available here.

- Enter one or more search terms.
- Available search fields are listed below
- Use Limits to restrict your search by search field, chromosome, and other criteria.

Update:	
January 5, 2005	Updated search terms
August 14, 2002	Add contig position tag [CTPOS]

Below are search examples and available search fields.

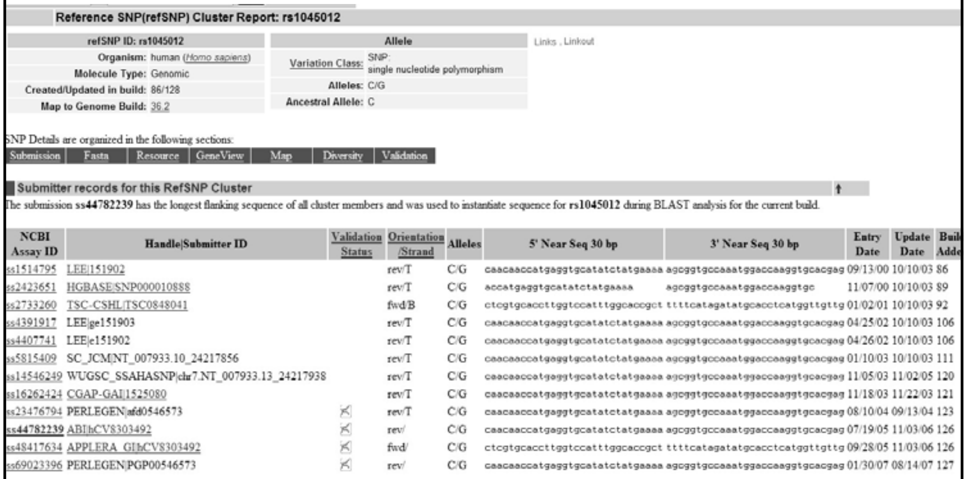
Search using wild-card(*), ranging(:), AND, OR, and NOT operators:	
Example	**Description**
BRC*[Gene Name]	Search SNPs on all genes with names starting with the letter 'BRC' (ie. BRCA1 and BRCA2)
1:5[HET]	Search SNPs with heterozygosity between 1 and 5 percent
coding nonsynonymous[FUNC] AND 1[CHR]	Search SNPs with function class 'coding nonsynonymous' located on chromosome 1
1[CHR] OR 2[CHR]	Search all SNPs on chromosome 1 or 2
1[CHR] OR 2[CHR] NOT unknown[METHOD]	Search all SNPs on chromosome 1 or 2 detected by all methods except 'unknown'.
1[WEIGHT] AND (1[CHR] OR 2[CHR]) NOT (unknown[METHOD] OR computed[METHOD])	Search all SNPs with weight 1 on chromosome 1 or 2 detected by all methods except 'unknown' or 'computed'.

Either the search fields or qualifiers (aliases) can be use for querying SNP (i.e. 103[CBID] is same as 103[Create Build ID]. Data type marked with an asterisk (*) indicates range searching is available.

Search Field	Qualifier	Type	Description
Allele	[ALLELE],[VARIATION], [VARI]	IUPAC	Observed allele(s) Example: N[ALLELE]
Chromosome	[CHR]	Textnum	Mapped chromosome number Available values [1-22,W-Z, and Un (unknown)] Example: 2[CHR] or X[CHR]
Base Position	[CHRPOS],[BPOS]	Integer*	Mapped chromosome position; use in conjunction with chromosome field [CHR] Example: 7[CHR] AND 88556398:88580839[CHRPOS]
Create Build ID	[CREATE_BUILD],[CBID]	Integer*	SNP create build ID Example: 103[CBID]
Publication Date	[CREATEDATE],[CDAT],[PDAT], [PUBDATE]	Date*	SNP create/publication date Use the format YYYY/MM/DD; month and day are optional. Example: "2005 07 13"[CDATE]
Function Class	[FXN_CLASS], [FUNC]	Text	Function Class: locus region intron coding nonsynonymous mrna utr coding synonymous reference exception splice site

http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp

dbSNP record for rs1045012



Reference SNP(refSNP) Cluster Report: rs1045012

refSNP ID: rs1045012	
Organism:	human (*Homo sapiens*)
Molecule Type:	Genomic
Created/Updated in build:	86/128
Map to Genome Build:	36.2

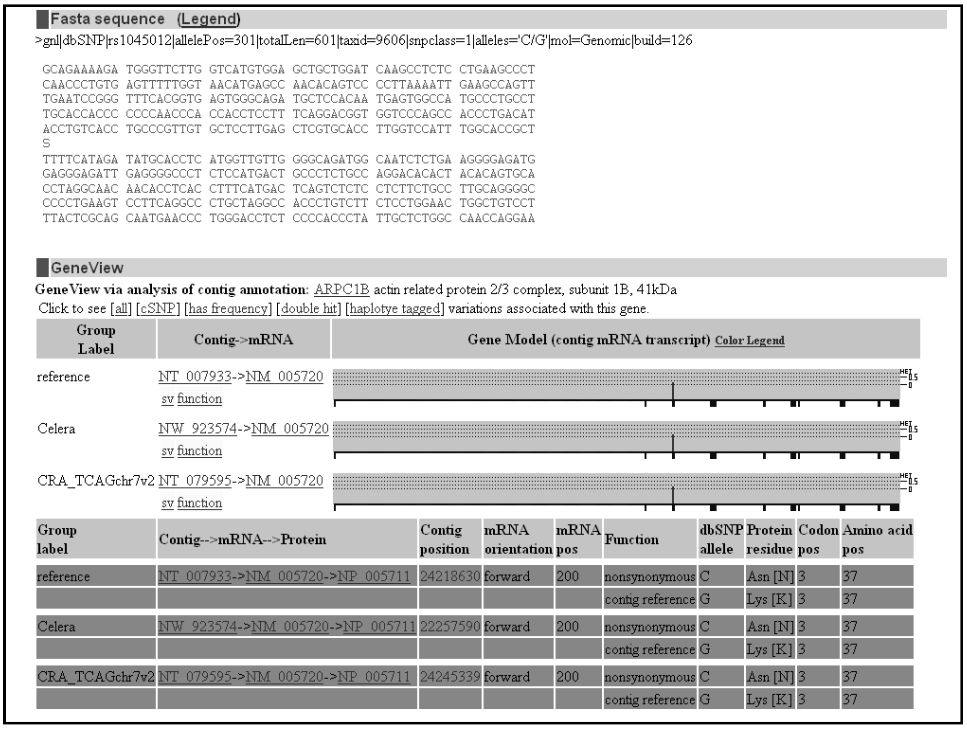
Allele	
Variation Class:	SNP: single nucleotide polymorphism
Alleles:	C/G
Ancestral Allele:	C

SNP Details are organized in the following sections:

Submitter records for this RefSNP Cluster

The submission ss**44782239** has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs1045012** during BLAST analysis for the current build.

NCBI Assay ID	Handle|Submitter ID	Validation Status	Orientation /Strand	Alleles	5' Near Seq 30 bp	3' Near Seq 30 bp	Entry Date	Update Date	Build Added
ss1514795	LEE|151902		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	09/13/00	10/10/03	86
ss2423651	HGBASE|SNP000010888		rev/T	C/G	accatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgc	11/07/00	10/10/03	89
ss2733260	TSC-CSHL|TSC0848041		fwd/B	C/G	ctcgtgcaccttggtccatttggcaccgct	ttttcatagatatgcacctcatggttgttg	01/02/01	10/10/03	92
ss4391917	LEE|ge151903		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	04/25/02	10/10/03	106
ss4407741	LEE|e151902		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	04/26/02	10/10/03	106
ss5815409	SC_JCM|NT_007933.10_24217856		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	01/10/03	10/10/03	111
ss14546249	WUGSC_SSAHASNP|chr7.NT_007933.13_24217938		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	11/05/03	11/02/05	120
ss16262424	CGAP-GAI|1525080		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	11/18/03	11/22/03	121
ss23476794	PERLEGEN|afd0546573	✓	rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	08/10/04	09/13/04	123
ss44782239	ABI|hCV8303492	✓	rev/	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	07/19/05	11/03/06	126
ss48417634	APPLERA_GI|hCV8303492	✓	fwd/	C/G	ctcgtgcaccttggtccatttggcaccgct	ttttcatagatatgcacctcatggttgttg	09/28/05	11/03/06	126
ss69023396	PERLEGEN|PGP00546573	✓	rev/	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	01/30/07	08/14/07	127



Fasta sequence (Legend)

```
>gnl|dbSNP|rs1045012|allelePos=301|totalLen=601|taxid=9606|snpclass=1|alleles='C/G'|mol=Genomic|build=126

 GCAGAAAAGA TGGGTTCTTG GTCATGTGGA GCTGCTGGAT CAAGCCTCTC CTGAAGCCCT
 CAACCCTGTG AGTTTTTGGT AACATGAGCC AACACAGTCC CCTTAAAATT GAAGCCAGTT
 TGAATCCGGG TTTCACGGTG AGTGGGCAGA TGCTCCACAA TGAGTGGCCA TGCCCTGCCT
 TGCACCACCC CCCCAACCCA CCACCTCCTT TCAGGACGGT GGTCCCAGCC ACCCTGACAT
 ACCTGTCACC TGCCCGTTGT GCTCCTTGAG CTCGTGCACC TTGGTCCATT TGGCACCGCT
 S
 TTTTCATAGA TATGCACCTC ATGGTTGTTG GGGCAGATGG CAATCTCTGA AGGGGAGATG
 GAGGGAGATT GAGGGGCCCT CTCCATGACT GCCCTCTGCC AGGACACACT ACACAGTGCA
 CCTAGGCAAC AACACCTCAC CTTTCATGAC TCAGTCTCTC CTCTTCTGCC TTGCAGGGGC
 CCCCTGAAGT CCTTCAGGCC CTGCTAGGCC ACCCTGTCTT CTCCTGGAAC TGGCTGTCCT
 TTACTCGCAG CAATGAACCC TGGGACCTCT CCCCACCCTA TTGCTCTGGC CAACCAGGAA
```

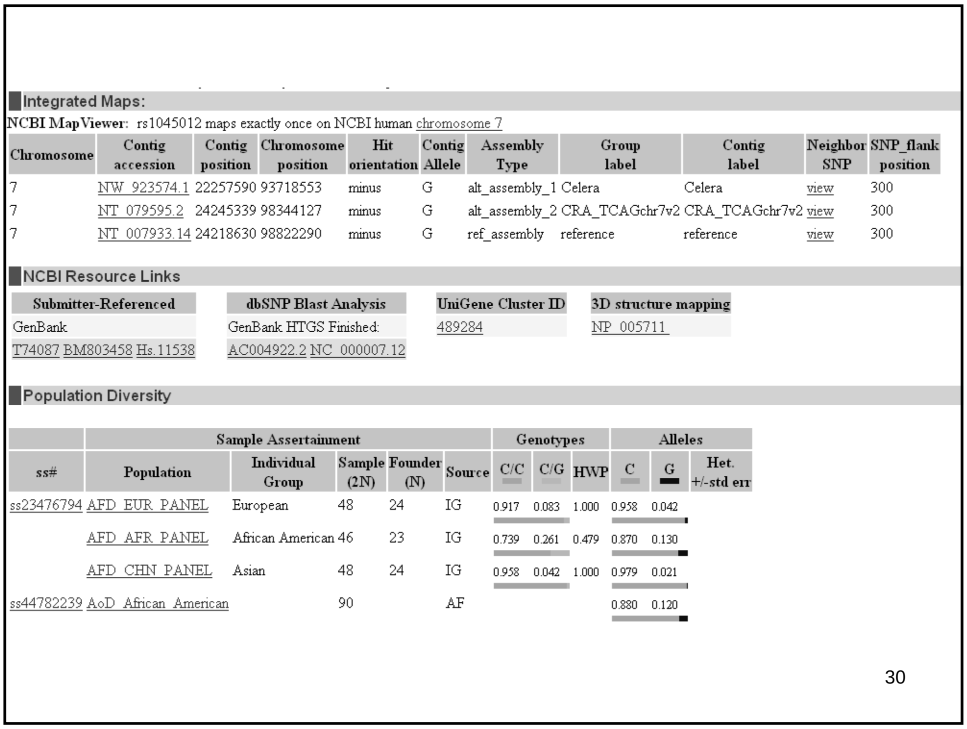
GeneView

GeneView via analysis of contig annotation: ARPC1B actin related protein 2/3 complex, subunit 1B, 41kDa

Click to see [all] [cSNP] [has frequency] [double hit] [haplotye tagged] variations associated with this gene.

Group Label	Contig->mRNA	Gene Model (contig mRNA transcript) Color Legend
reference	NT_007933->NM_005720 sv function	
Celera	NW_923574->NM_005720 sv function	
CRA_TCAGchr7v2	NT_079595->NM_005720 sv function	

Group label	Contig-->mRNA-->Protein	Contig position	mRNA orientation	mRNA pos	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
reference	NT_007933->NM_005720->NP_005711	24218630	forward	200	nonsynonymous	C	Asn [N]	3	37
					contig reference	G	Lys [K]	3	37
Celera	NW_923574->NM_005720->NP_005711	22257590	forward	200	nonsynonymous	C	Asn [N]	3	37
					contig reference	G	Lys [K]	3	37
CRA_TCAGchr7v2	NT_079595->NM_005720->NP_005711	24245339	forward	200	nonsynonymous	C	Asn [N]	3	37
					contig reference	G	Lys [K]	3	37



Integrated Maps:

NCBI MapViewer: rs1045012 maps exactly once on NCBI human chromosome 7

Chromosome	Contig accession	Contig position	Chromosome position	Hit orientation	Contig Allele	Assembly Type	Group label	Contig label	Neighbor SNP	SNP_flank position
7	NW_923574.1	22257590	93718553	minus	G	alt_assembly_1	Celera	Celera	view	300
7	NT_079595.2	24245339	98344127	minus	G	alt_assembly_2	CRA_TCAGchr7v2	CRA_TCAGchr7v2	view	300
7	NT_007933.14	24218630	98822290	minus	G	ref_assembly	reference	reference	view	300

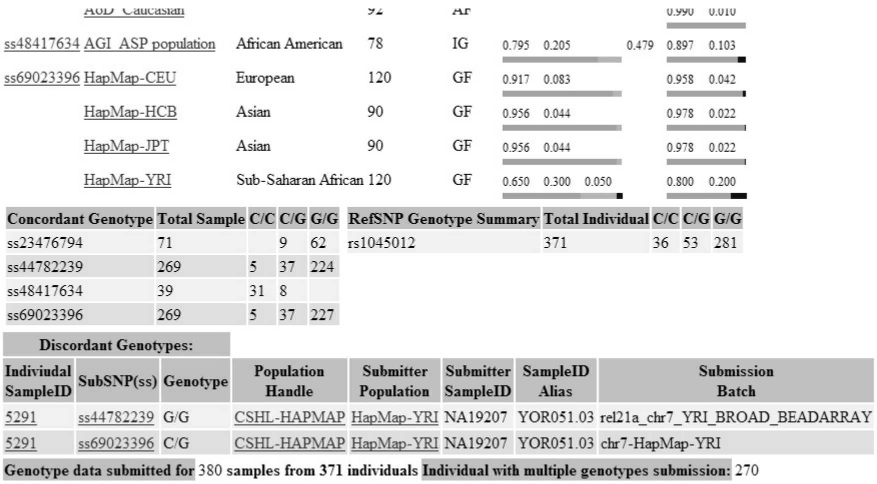
NCBI Resource Links

Submitter-Referenced	dbSNP Blast Analysis	UniGene Cluster ID	3D structure mapping
GenBank	GenBank HTGS Finished:	489284	NP_005711
T74087 BM803458 Hs.11538	AC004922.2 NC_000007.12		

Population Diversity

	Sample Assertainment					Genotypes			Alleles		
ss#	Population	Individual Group	Sample (2N)	Founder (N)	Source	C/C	C/G	HWP	C	G	Het. +/-std err
ss23476794	AFD_EUR_PANEL	European	48	24	IG	0.917	0.083	1.000	0.958	0.042	
	AFD_AFR_PANEL	African American	46	23	IG	0.739	0.261	0.479	0.870	0.130	
	AFD_CHN_PANEL	Asian	48	24	IG	0.958	0.042	1.000	0.979	0.021	
ss44782239	AoD_African_American		90		AF				0.880	0.120	

30



	AoD_Caucasian		92	AF				0.990	0.010
ss48417634	AGI_ASP population	African American	78	IG	0.795	0.205	0.479	0.897	0.103
ss69023396	HapMap-CEU	European	120	GF	0.917	0.083		0.958	0.042
	HapMap-HCB	Asian	90	GF	0.956	0.044		0.978	0.022
	HapMap-JPT	Asian	90	GF	0.956	0.044		0.978	0.022
	HapMap-YRI	Sub-Saharan African	120	GF	0.650	0.300 0.050		0.800	0.200

Concordant Genotype	Total Sample	C/C	C/G	G/G
ss23476794	71		9	62
ss44782239	269	5	37	224
ss48417634	39	31	8	
ss69023396	269	5	37	227

RefSNP Genotype Summary	Total Individual	C/C	C/G	G/G
rs1045012	371	36	53	281

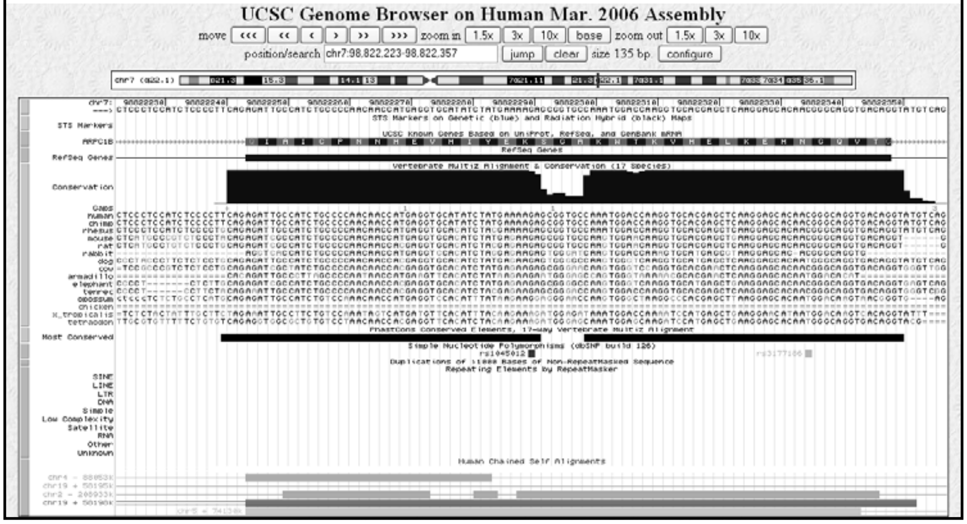
Discordant Genotypes:

Indiviudal SampleID	SubSNP(ss)	Genotype	Population Handle	Submitter Population	Submitter SampleID	SampleID Alias	Submission Batch
5291	ss44782239	G/G	CSHL-HAPMAP	HapMap-YRI	NA19207	YOR051.03	rel21a_chr7_YRI_BROAD_BEADARRAY
5291	ss69023396	C/G	CSHL-HAPMAP	HapMap-YRI	NA19207	YOR051.03	chr7-HapMap-YRI

Genotype data submitted for 380 **samples from 371 individuals** **Individual with multiple genotypes submission:** 270

Viewing SNPs in Browsers

NCBI Ensembl UCSC



UCSC Genome Browser on Human Mar. 2006 Assembly

position/search chr7:98,822,223-98,822,357 size 135 bp

move
‹‹‹
‹‹
‹
›
››
›››
zoom in
1.5x
3x
10x
base
zoom out
1.5x
3x
10x
jump
clear
configure
chr7:
STS Markers
ARPC1B
RefSeq Genes
Conservation
Gaps
human
chimp
rhesus
mouse
rat
rabbit
dog
cow
armadillo
elephant
tenrec
opossum
chicken
X_tropicalis
tetraodon
Most Conserved
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
RefSeq Genes
Vertebrate Multiz Alignment & Conservation (17 Species)
PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment
Simple Nucleotide Polymorphisms (dbSNP build 126)
rs1045012
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Repeating Elements by RepeatMasker
Human Chained Self Alignments

Overview of Topics

- Review of genetic variation discovery
- Database of SNPs, dbSNP
- Other types of genetic variation
- Medical sequencing
- Next-generation sequencing and SNPs
- Targeted Genomic Selection

Other Types of Sequence Variation

- Deletion/Insertion Polymorphisms (DIPs)
  - Also called indels, sizes from 1base to ~1kb
  - More difficult to detect and automatically type
  - Occur at less frequent intervals; about 8 times less frequent compared to SNPs
    - 2.1M DIPs and 9.3M SNPs
    - More difficult to cluster, e.g. rs34505627 and rs10581774:

```
atttatttatttattt reference
attt----atttattt rs10581774
atttattta----ttt rs34505627
```

- Structural Variation
- Copy Number Variation

Definition of Terms: Larger Scale Variation

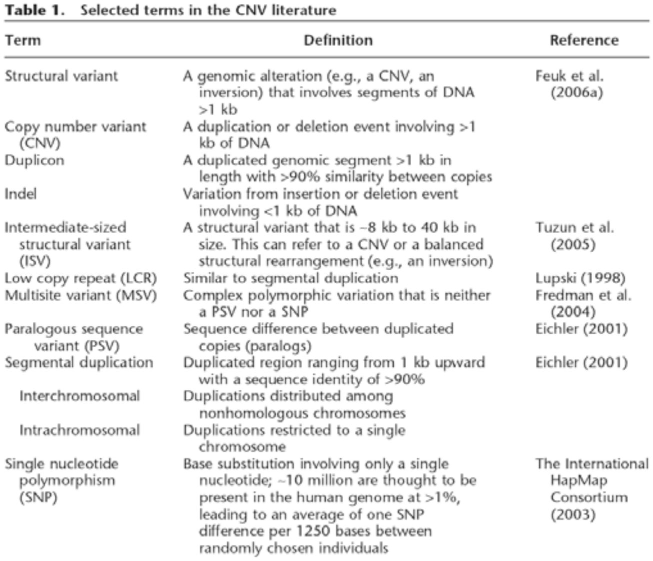


Table 1. Selected terms in the CNV literature

Term	Definition	Reference
Structural variant	A genomic alteration (e.g., a CNV, an inversion) that involves segments of DNA >1 kb	Feuk et al. (2006a)
Copy number variant (CNV)	A duplication or deletion event involving >1 kb of DNA	
Duplicon	A duplicated genomic segment >1 kb in length with >90% similarity between copies	
Indel	Variation from insertion or deletion event involving <1 kb of DNA	
Intermediate-sized structural variant (ISV)	A structural variant that is ~8 kb to 40 kb in size. This can refer to a CNV or a balanced structural rearrangement (e.g., an inversion)	Tuzun et al. (2005)
Low copy repeat (LCR)	Similar to segmental duplication	Lupski (1998)
Multisite variant (MSV)	Complex polymorphic variation that is neither a PSV nor a SNP	Fredman et al. (2004)
Paralogous sequence variant (PSV)	Sequence difference between duplicated copies (paralogs)	Eichler (2001)
Segmental duplication	Duplicated region ranging from 1 kb upward with a sequence identity of >90%	Eichler (2001)
Interchromosomal	Duplications distributed among nonhomologous chromosomes	
Intrachromosomal	Duplications restricted to a single chromosome	
Single nucleotide polymorphism (SNP)	Base substitution involving only a single nucleotide; ~10 million are thought to be present in the human genome at >1%, leading to an average of one SNP difference per 1250 bases between randomly chosen individuals	The International HapMap Consortium (2003)

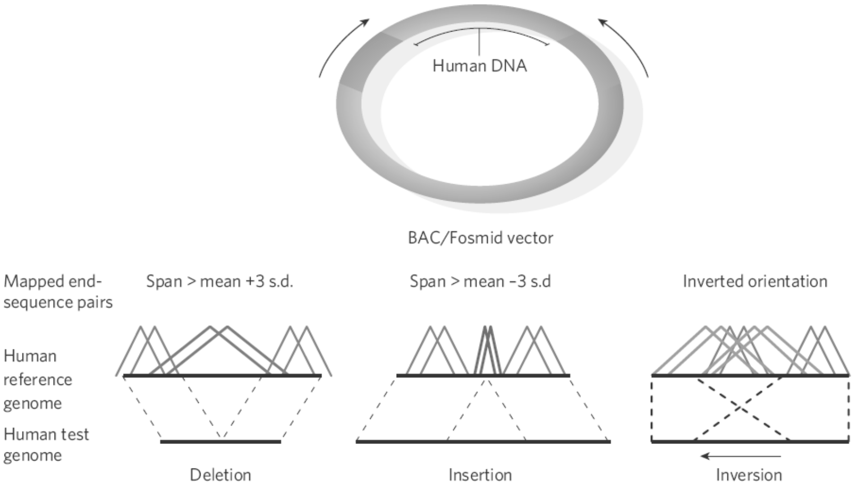
Genome Res. 2006 16: 949-961

35

Human Genome Structural Variation Project

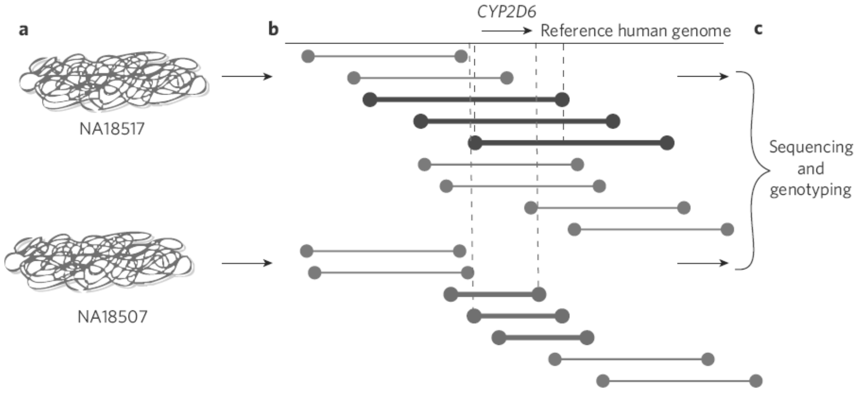
- NHGRI funded initiative
- A sequence-based survey of human structural variation aims to characterize common structural variants that are larger than (>5 kb)
- Types include multi-kilobase deletions, insertions, inversions, translocations, and duplications
- The approach entails sequencing the ends of fosmids and BACs from multiple individuals

36



Nature. 2007 May 10;447(7141):161-5.

- Sequence from 1M fosmid ends (2M sequence reads) covers the genome to about 12X template coverage
- Or about 6X coverage across each of an individuals 2 haploids
- This is enough coverage to detect deletions and insertion as small as about 5kb



Nature. 2007 May 10;447(7141):161-5.

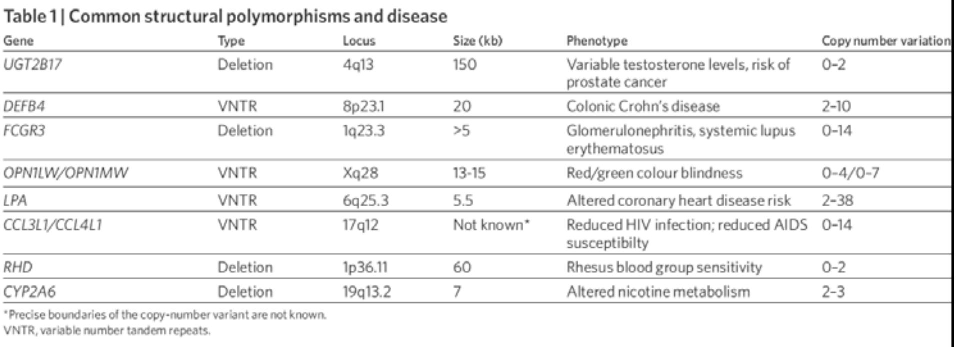


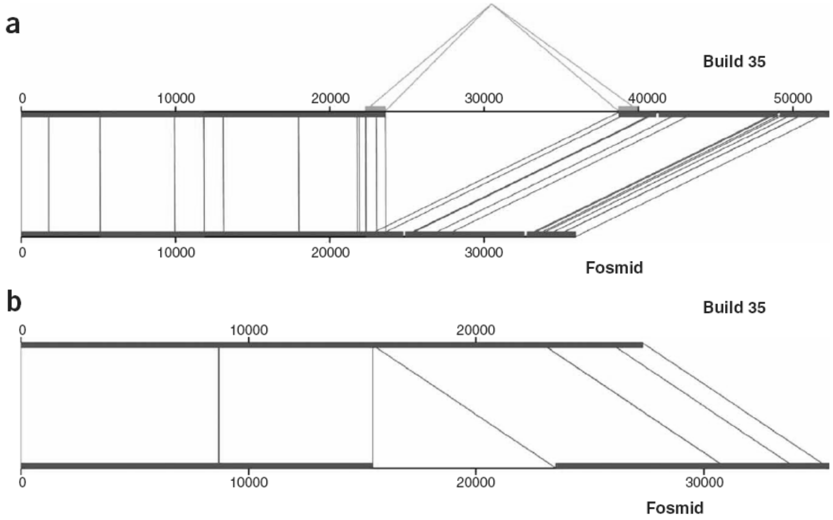
Table 1 | Common structural polymorphisms and disease

Gene	Type	Locus	Size (kb)	Phenotype	Copy number variation
UGT2B17	Deletion	4q13	150	Variable testosterone levels, risk of prostate cancer	0–2
DEFB4	VNTR	8p23.1	20	Colonic Crohn's disease	2–10
FCGR3	Deletion	1q23.3	>5	Glomerulonephritis, systemic lupus erythematosus	0–14
OPN1LW/OPN1MW	VNTR	Xq28	13-15	Red/green colour blindness	0–4/0–7
LPA	VNTR	6q25.3	5.5	Altered coronary heart disease risk	2–38
CCL3L1/CCL4L1	VNTR	17q12	Not known*	Reduced HIV infection; reduced AIDS susceptibilty	0–14
RHD	Deletion	1p36.11	60	Rhesus blood group sensitivity	0–2
CYP2A6	Deletion	19q13.2	7	Altered nicotine metabolism	2–3

*Precise boundaries of the copy-number variant are not known.
VNTR, variable number tandem repeats.

Nature. 2007 May 10;447(7141):161-5.

Sequence level identification of deletion and insertion events



Nature Genetics **37**, 727 - 732 (2005)

Structural Variation Project Goals

- Generate fosmid and BAC end sequence data for up to 48 HapMap individuals
- Sequence for 9 individuals are available
- Twelve more are “ongoing”
- Mine the data for common and rare structural variants
- Mine the trace data for SNPs and DIPs

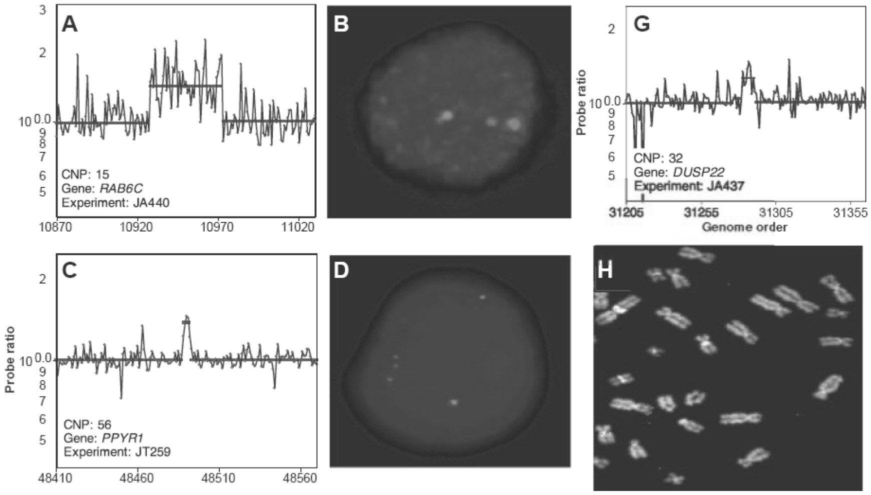
http://www.genome.gov/25521748

Copy Number Variation

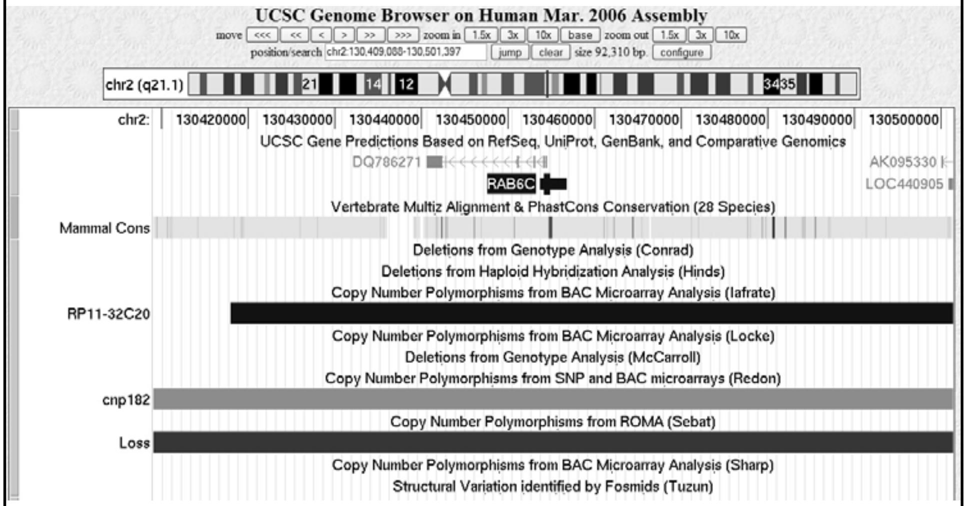
- This is structural variation, however the methods used to detect CNVs do not give precise local structural information

- Typically detected using an array-based technology, e.g.
  - SNP genotyping chips
  - Oligonucleotide arrays

Copy number variation detected using representational oligonucleotide microarray analysis (ROMA)



Science. 2004 July 23;305(5683):525-8



UCSC Genome Browser on Human Mar. 2006 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr2:130,409,088-130,501,397 jump clear size 92,310 bp. configure
chr2 (q21.1)
21
14
12
34
35
chr2:
130420000
130430000
130440000
130450000
130460000
130470000
130480000
130490000
130500000

UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics

Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)

Deletions from Genotype Analysis (Conrad)

Deletions from Haploid Hybridization Analysis (Hinds)

Copy Number Polymorphisms from BAC Microarray Analysis (Iafrate)

Copy Number Polymorphisms from BAC Microarray Analysis (Locke)

Deletions from Genotype Analysis (McCarroll)

Copy Number Polymorphisms from SNP and BAC microarrays (Redon)

Copy Number Polymorphisms from ROMA (Sebat)

Copy Number Polymorphisms from BAC Microarray Analysis (Sharp)

Structural Variation identified by Fosmids (Tuzun)

Future of CNV detection

- New SNP chips are being designed to include more features to detect CNVs at a higher resolution across the genome
- These new chips will be applied to many more samples

45

Overview of Topics

- Review of genetic variation discovery
- Database of SNPs, dbSNP
- Other types of genetic variation
- Medical sequencing
- Next-generation sequencing and SNPs
- Targeted Genomic Selection

46

Medical Sequencing Project Initiatives

- Mapped Autosomal Mendelian Disorders
- Allelic Spectrum in Common Disease

  http://www.genome.gov/20019648
- Tumor Sequencing Project

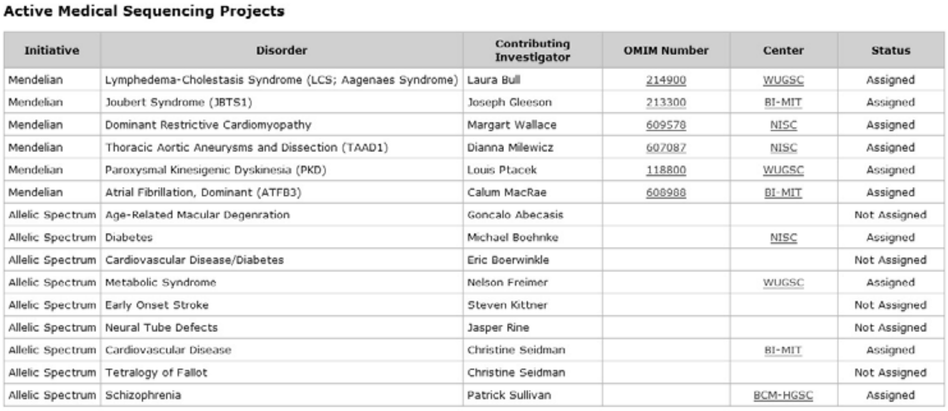
  http://www.genome.gov/19517442
- The Cancer Genome Atlas Project
  - NCI GRAND ROUNDS Lecture by Dr. Collins

  http://videocast.nih.gov/Summary.asp?File=14383

  http://cancergenome.nih.gov/about/index.asp

Mendelian Initiative:
- mapped Mendelian disorders to intervals of about 10 Mb or less

Allelic Spectrum Initiative:
- sequencing genes implicated in common disorders in large, well-phenotyped cohorts



Active Medical Sequencing Projects

Initiative	Disorder	Contributing Investigator	OMIM Number	Center	Status
Mendelian	Lymphedema-Cholestasis Syndrome (LCS; Aagenaes Syndrome)	Laura Bull	214900	WUGSC	Assigned
Mendelian	Joubert Syndrome (JBTS1)	Joseph Gleeson	213300	BI-MIT	Assigned
Mendelian	Dominant Restrictive Cardiomyopathy	Margart Wallace	609578	NISC	Assigned
Mendelian	Thoracic Aortic Aneurysms and Dissection (TAAD1)	Dianna Milewicz	607087	NISC	Assigned
Mendelian	Paroxysmal Kinesigenic Dyskinesia (PKD)	Louis Ptacek	118800	WUGSC	Assigned
Mendelian	Atrial Fibrillation, Dominant (ATFB3)	Calum MacRae	608988	BI-MIT	Assigned
Allelic Spectrum	Age-Related Macular Degenration	Goncalo Abecasis			Not Assigned
Allelic Spectrum	Diabetes	Michael Boehnke		NISC	Assigned
Allelic Spectrum	Cardiovascular Disease/Diabetes	Eric Boerwinkle			Not Assigned
Allelic Spectrum	Metabolic Syndrome	Nelson Freimer		WUGSC	Assigned
Allelic Spectrum	Early Onset Stroke	Steven Kittner			Not Assigned
Allelic Spectrum	Neural Tube Defects	Jasper Rine			Not Assigned
Allelic Spectrum	Cardiovascular Disease	Christine Seidman		BI-MIT	Assigned
Allelic Spectrum	Tetralogy of Fallot	Christine Seidman			Not Assigned
Allelic Spectrum	Schizophrenia	Patrick Sullivan		BCM-HGSC	Assigned

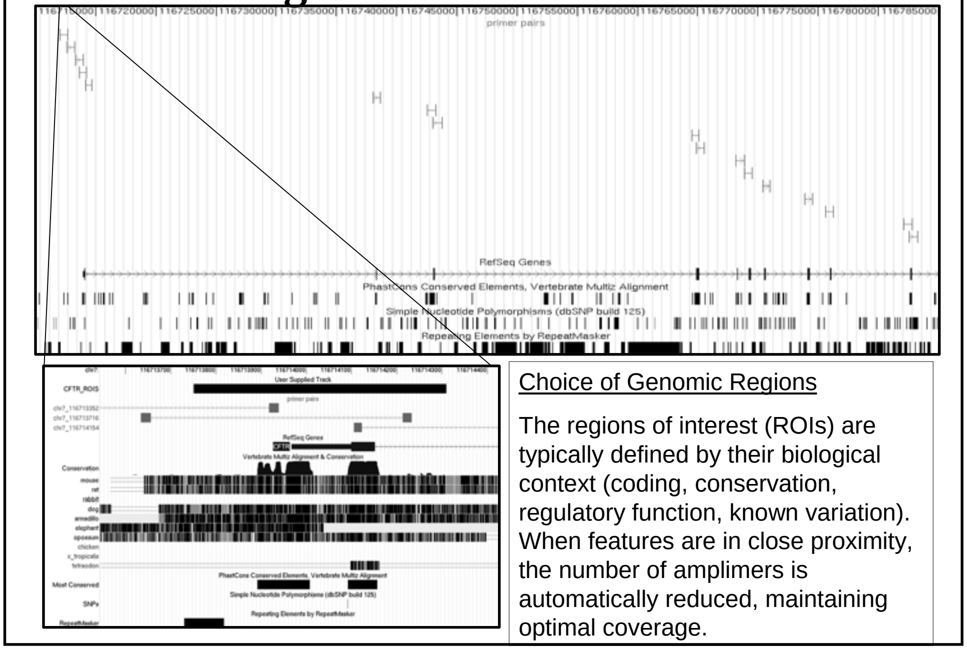
http://www.genome.gov/20019648

Medical Sequencing

- This is accomplished using PCR amplification of selected targets followed by Sanger sequencing
  - Regions of interest (ROI) are defined, e.g. all coding exons in a suspected disease gene
  - PCR primer pairs designed to cover ROIs
  - PCR amplification and sequencing
  - Sequence variant detection

49

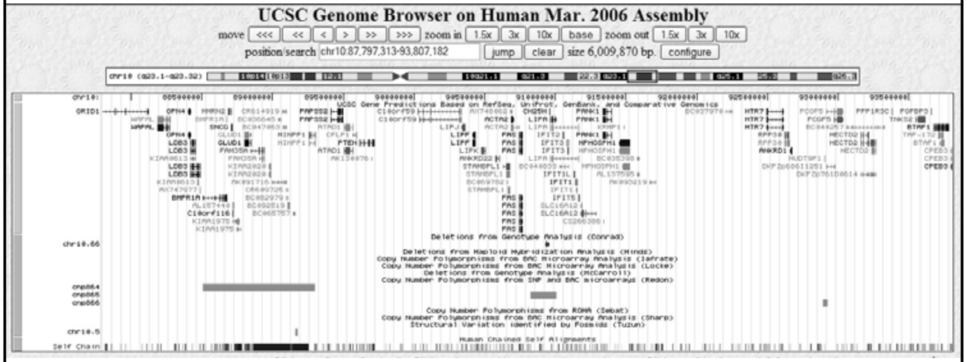
Primer Design



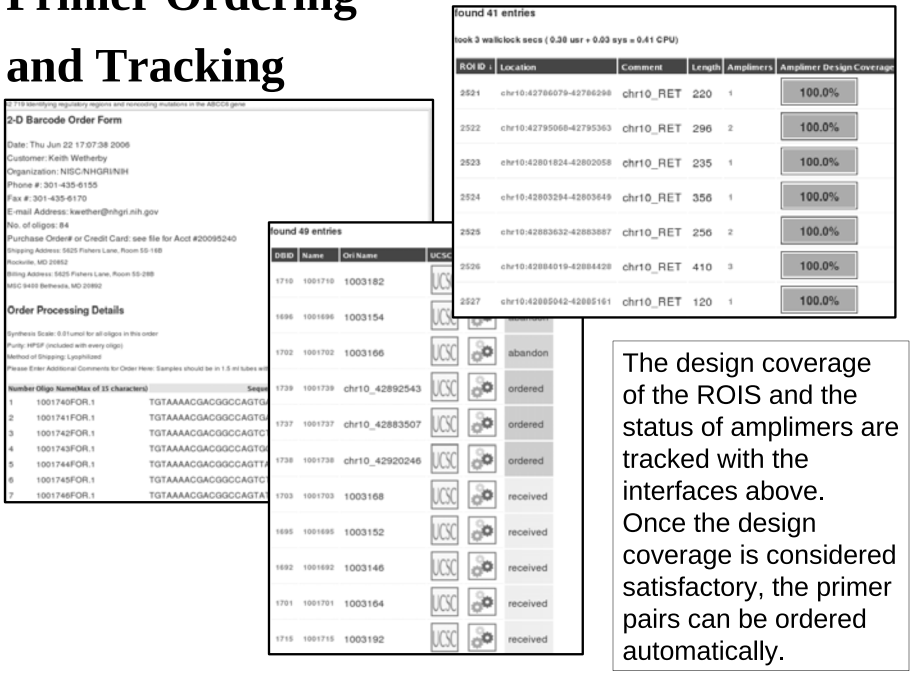
Choice of Genomic Regions

The regions of interest (ROIs) are typically defined by their biological context (coding, conservation, regulatory function, known variation). When features are in close proximity, the number of amplimers is automatically reduced, maintaining optimal coverage.

Watch out for segmental duplications or CNVs


UCSC Genome Browser on Human Mar. 2006 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr10:87,797,313-93,807,182 jump clear size 6,009,870 bp. configure
chr10 (q23.1-q23.32)
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
Deletions from Genotype Analysis (Conrad)
Deletions from Haploid Hybridization Analysis (Hinds)
Copy Number Polymorphisms from BAC Microarray Analysis (Iafrate)
Copy Number Polymorphisms from BAC Microarray Analysis (Locke)
Deletions from Genotype Analysis (McCarroll)
Copy Number Polymorphisms from SNP and BAC microarrays (Redon)
Copy Number Polymorphisms from ROMA (Sebat)
Copy Number Polymorphisms from BAC Microarray Analysis (Sharp)
Structural Variation identified by Fosmids (Tuzun)
Human Chained Self Alignments
Self Chain

Primer Ordering and Tracking



2.719 identifying regulatory regions and noncoding mutations in the ABCC6 gene

2-D Barcode Order Form

Date: Thu Jun 22 17:07:38 2006
Customer: Keith Wetherby
Organization: NISC/NHGRI/NIH
Phone #: 301-435-6155
Fax #: 301-435-6170
E-mail Address: kwether@nhgri.nih.gov
No. of oligos: 84
Purchase Order# or Credit Card: see file for Acct #20095240
Shipping Address: 5625 Fishers Lane, Room 5S-16B
Rockville, MD 20852
Billing Address: 5625 Fishers Lane, Room 5S-28B
MSC 9400 Bethesda, MD 20892

Order Processing Details

Synthesis Scale: 0.01umol for all oligos in this order
Purity: HPSF (included with every oligo)
Method of Shipping: Lyophilized
Please Enter Additional Comments for Order Here: Samples should be in 1.5 ml tubes wit

Number	Oligo Name(Max of 15 characters)	Seque
1	1001740FOR.1	TGTAAAACGACGGCCAGTGA
2	1001741FOR.1	TGTAAAACGACGGCCAGTGA
3	1001742FOR.1	TGTAAAACGACGGCCAGTCT
4	1001743FOR.1	TGTAAAACGACGGCCAGTGC
5	1001744FOR.1	TGTAAAACGACGGCCAGTTA
6	1001745FOR.1	TGTAAAACGACGGCCAGTCT
7	1001746FOR.1	TGTAAAACGACGGCCAGTAT

found 49 entries

DBID	Name	Ori Name	UCSC		
1710	1001710	1003182	UCSC		
1696	1001696	1003154	UCSC	⚙	abandon
1702	1001702	1003166	UCSC	⚙	abandon
1739	1001739	chr10_42892543	UCSC	⚙	ordered
1737	1001737	chr10_42883507	UCSC	⚙	ordered
1738	1001738	chr10_42920246	UCSC	⚙	ordered
1703	1001703	1003168	UCSC	⚙	received
1695	1001695	1003152	UCSC	⚙	received
1692	1001692	1003146	UCSC	⚙	received
1701	1001701	1003164	UCSC	⚙	received
1715	1001715	1003192	UCSC	⚙	received

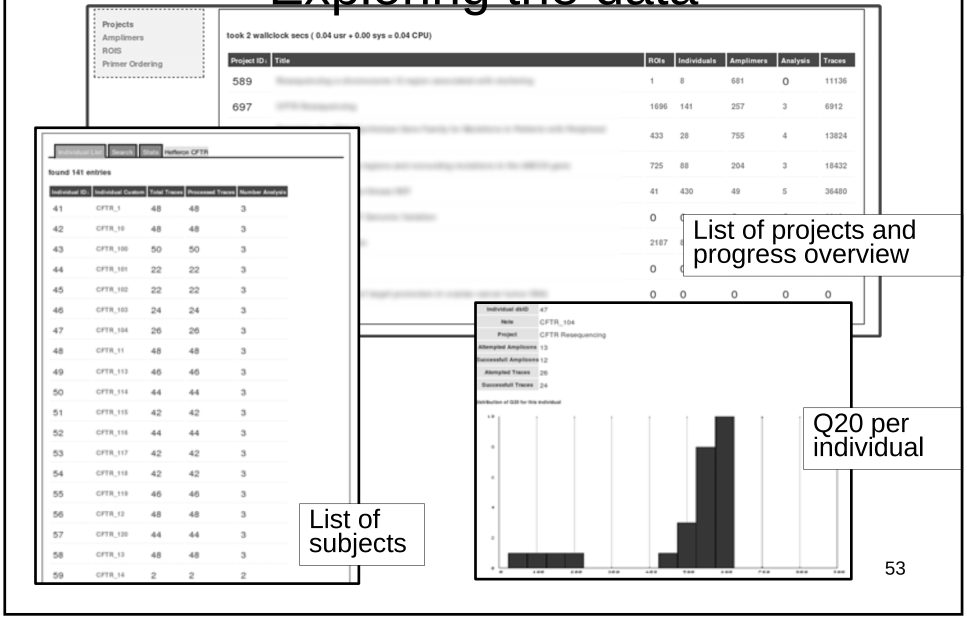
found 41 entries

took 3 wallclock secs (0.38 usr + 0.03 sys = 0.41 CPU)

ROI ID ↓	Location	Comment	Length	Amplimers	Amplimer Design Coverage
2521	chr10:42786079-42786298	chr10_RET	220	1	100.0%
2522	chr10:42795068-42795363	chr10_RET	296	2	100.0%
2523	chr10:42801824-42802058	chr10_RET	235	1	100.0%
2524	chr10:42803294-42803649	chr10_RET	356	1	100.0%
2525	chr10:42883632-42883887	chr10_RET	256	2	100.0%
2526	chr10:42884019-42884428	chr10_RET	410	3	100.0%
2527	chr10:42885042-42885161	chr10_RET	120	1	100.0%

The design coverage of the ROIS and the status of amplimers are tracked with the interfaces above. Once the design coverage is considered satisfactory, the primer pairs can be ordered automatically.

Exploring the data



- Projects
- Amplimers
- ROIS
- Primer Ordering

took 2 wallclock secs (0.04 usr + 0.00 sys = 0.04 CPU)

Project ID↓	Title	ROIs	Individuals	Amplimers	Analysis	Traces
589	[illegible]	1	8	681	0	11136
697	[illegible]	1696	141	257	3	6912
	[illegible]	433	28	755	4	13824
	[illegible]	725	88	204	3	18432
	[illegible]	41	430	49	5	36480
	[illegible]	0				
	[illegible]	2187				
		0				
	[illegible]	0	0	0	0	0

List of projects and progress overview

Individual List | Search | Stats | Hefferon CFTR

found 141 entries

Individual ID↓	Individual Custom	Total Traces	Processed Traces	Number Analysis
41	CFTR_1	48	48	3
42	CFTR_10	48	48	3
43	CFTR_100	50	50	3
44	CFTR_101	22	22	3
45	CFTR_102	22	22	3
46	CFTR_103	24	24	3
47	CFTR_104	26	26	3
48	CFTR_11	48	48	3
49	CFTR_113	46	46	3
50	CFTR_114	44	44	3
51	CFTR_115	42	42	3
52	CFTR_116	44	44	3
53	CFTR_117	42	42	3
54	CFTR_118	42	42	3
55	CFTR_119	46	46	3
56	CFTR_12	48	48	3
57	CFTR_120	44	44	3
58	CFTR_13	48	48	3
59	CFTR_14	2	2	2

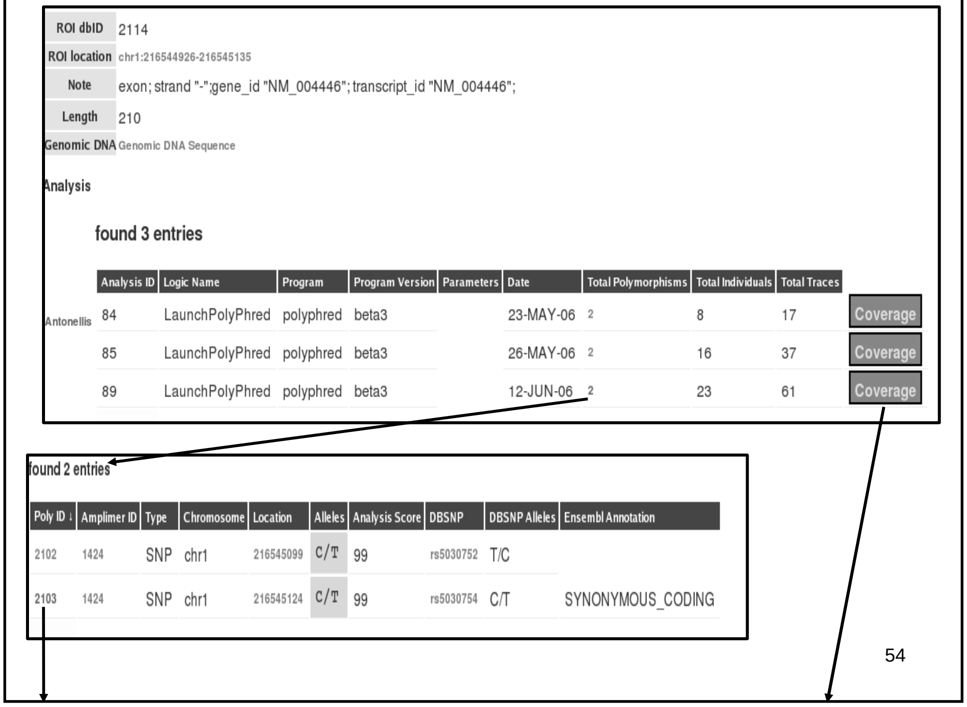
List of subjects

Individual dbID	47
Note	CFTR_104
Project	CFTR Resequencing
Attempted Amplicons	13
Successfull Amplicons	12
Atempted Traces	26
Successfull Traces	24

Distribution of Q20 for this individual

Q20 per individual

53



ROI dbID	2114
ROI location	chr1:216544926-216545135
Note	exon; strand "-";gene_id "NM_004446"; transcript_id "NM_004446";
Length	210
Genomic DNA	Genomic DNA Sequence

Analysis

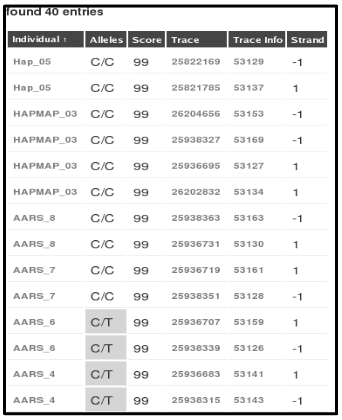
found 3 entries

	Analysis ID	Logic Name	Program	Program Version	Parameters	Date	Total Polymorphisms	Total Individuals	Total Traces	
Antonellis	84	LaunchPolyPhred	polyphred	beta3		23-MAY-06	2	8	17	Coverage
	85	LaunchPolyPhred	polyphred	beta3		26-MAY-06	2	16	37	Coverage
	89	LaunchPolyPhred	polyphred	beta3		12-JUN-06	2	23	61	Coverage

found 2 entries

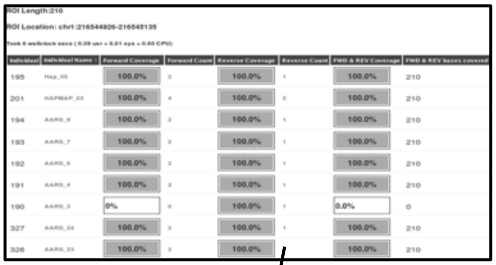
Poly ID ↓	Amplimer ID	Type	Chromosome	Location	Alleles	Analysis Score	DBSNP	DBSNP Alleles	Ensembl Annotation
2102	1424	SNP	chr1	216545099	C/T	99	rs5030752	T/C	
2103	1424	SNP	chr1	216545124	C/T	99	rs5030754	C/T	SYNONYMOUS_CODING

54



found 40 entries

Individual ↑	Alleles	Score	Trace	Trace Info	Strand
Hap_05	C/C	99	25822169	53129	-1
Hap_05	C/C	99	25821785	53137	1
HAPMAP_03	C/C	99	26204656	53153	-1
HAPMAP_03	C/C	99	25938327	53169	-1
HAPMAP_03	C/C	99	25936695	53127	1
HAPMAP_03	C/C	99	26202832	53134	1
AARS_8	C/C	99	25938363	53163	-1
AARS_8	C/C	99	25936731	53130	1
AARS_7	C/C	99	25936719	53161	1
AARS_7	C/C	99	25938351	53128	-1
AARS_6	C/T	99	25936707	53159	1
AARS_6	C/T	99	25938339	53126	-1
AARS_4	C/T	99	25936683	53141	1
AARS_4	C/T	99	25938315	53143	-1

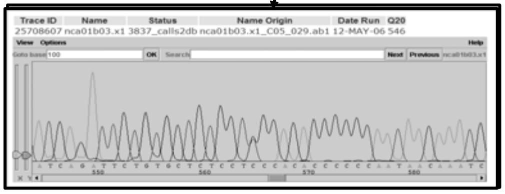


ROI Length:210

ROI Location: chr1:216544926-216545135

Took 6 wallclock secs (0.39 usr + 0.01 sys = 0.40 CPU)

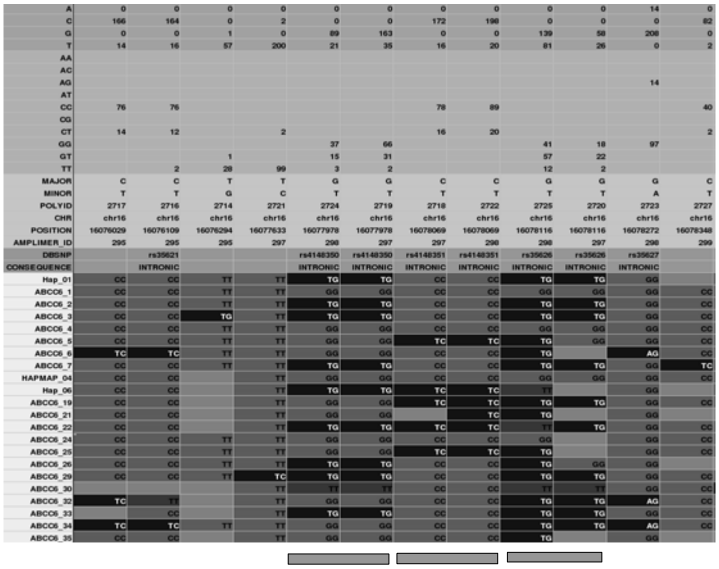
Individual	Individual Name ↑	Forward Coverage	Forward Count	Reverse Coverage	Reverse Count	FWD & REV Coverage	FWD & REV bases covered
195	Hap_05	100.0%	2	100.0%	1	100.0%	210
201	HAPMAP_03	100.0%	4	100.0%	2	100.0%	210
194	AARS_8	100.0%	2	100.0%	1	100.0%	210
193	AARS_7	100.0%	2	100.0%	1	100.0%	210
192	AARS_6	100.0%	2	100.0%	1	100.0%	210
191	AARS_4	100.0%	2	100.0%	1	100.0%	210
190	AARS_3	0%	0	100.0%	1	0.0%	0
327	AARS_24	100.0%	2	100.0%	1	100.0%	210
326	AARS_23	100.0%	2	100.0%	1	100.0%	210



Trace ID	Name	Status	Name Origin	Date Run	Q20
25708607	nca01b03.x1	3837_calls2db	nca01b03.x1_C05_029.ab1	12-MAY-06	546

View
Options
Help
Goto base 100
OK
Search
Next
Previous
nca01b03.x1
A T C A G A T T C T G T G C T C T C C T C C C A C A C C C C C C A A T A A C A A A T C
550
560
570
580

The system keeps track of analysis performed on the data and coverage attained for each ROI. It also allows a user to browse the detected genotypes.

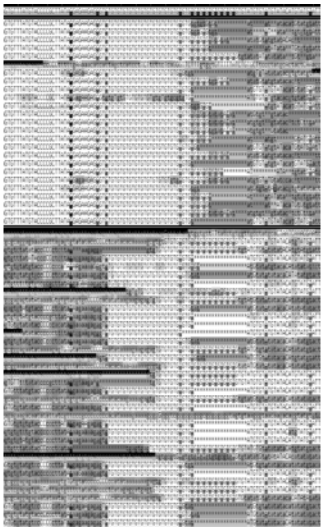


A	0	0	0	0	0	0	0	0	0	0	14	0
C	166	164	0	2	0	0	172	198	0	0	0	82
G	0	0	1	0	89	163	0	0	139	58	208	0
T	14	16	57	200	21	35	16	20	81	26	0	2
AA												
AC												
AG											14	
AT												
CC	76	76					78	89				40
CG												
CT	14	12		2			16	20				2
GG					37	66			41	18	97	
GT			1		15	31			57	22		
TT		2	28	99	3	2			12	2		
MAJOR	C	C	T	T	G	G	C	C	G	G	G	C
MINOR	T	T	G	C	T	T	T	T	T	T	A	T
POLYID	2717	2716	2714	2721	2724	2719	2718	2722	2725	2720	2723	2727
CHR	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16
POSITION	16076029	16076109	16076294	16077633	16077978	16077978	16078069	16078069	16078116	16078116	16078272	16078348
AMPLIMER_ID	295	295	295	297	298	297	297	298	298	297	298	299
DBSNP		rs35621			rs4148350	rs4148350	rs4148351	rs4148351	rs35626	rs35626	rs35627	
CONSEQUENCE		INTRONIC			INTRONIC	INTRONIC	INTRONIC	INTRONIC	INTRONIC	INTRONIC	INTRONIC	
Hap_01	CC	CC	TT	TT	TG	TG	CC	CC	TG	TG	GG	
ABCC6_1	CC	CC	TT	TT	GG	GG	CC	CC	GG	GG	GG	CC
ABCC6_2	CC	CC	TT	TT	TG	TG	CC	CC	TG	TG	GG	CC
ABCC6_3	CC	CC	TG	TT	TG	TG	CC	CC	TG	TG	GG	CC
ABCC6_4	CC	CC	TT	TT	GG	GG	CC	CC	GG	GG	GG	CC
ABCC6_5	CC	CC	TT	TT	GG	GG	TC	TC	TG	GG	GG	CC
ABCC6_6	TC	TC	TT	TT	GG	GG	CC	CC	TG		AG	CC
ABCC6_7	CC	CC	TT	TT	TG	TG	CC	CC	TG	TG	GG	TC
HAPMAP_04	CC	CC		TT	GG	GG	CC	CC	GG	GG	GG	CC
Hap_06	CC	CC		TT	TG	TG	TC	TC	TT		GG	
ABCC6_19	CC	CC		TT	GG	GG	TC	TC	TG	TG	GG	CC
ABCC6_21	CC	CC		TT	GG	GG		TC	TG		GG	
ABCC6_22	CC	CC		TT	TG	TG	TC	TC	TT	TG	GG	CC
ABCC6_24	CC	CC	TT	TT	GG	GG	CC	CC	GG		GG	CC
ABCC6_25	CC	CC	TT	TT	GG	GG	TC	TC	TG		GG	CC
ABCC6_26	CC	CC	TT	TT	TG	TG	CC	CC	TG	GG	GG	
ABCC6_29	CC	CC	TT	TC	TG	TG	CC	CC	TG	TG	GG	CC
ABCC6_30				TT	TT	TT	CC	CC	TT	TT	GG	CC
ABCC6_32	TC	TT		TT	GG	GG	CC	CC	TG	TG	AG	CC
ABCC6_33		CC		TT	TG	TG	CC	CC	TG	TG	GG	CC
ABCC6_34	TC	TC	TT	TT	GG	GG	CC	CC	TG	TG	AG	CC
ABCC6_35	CC	CC		TT	GG	GG	CC	CC	TG		GG	

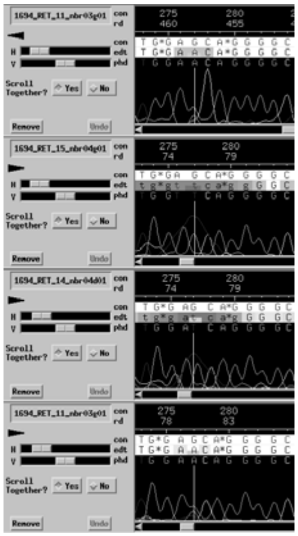
We are developing interfaces that allow exploring the results and identify interesting results as well flag problems.

Three examples of same SNP detected in overlapping amplimers. This information is used to assess accuracy of the detection.

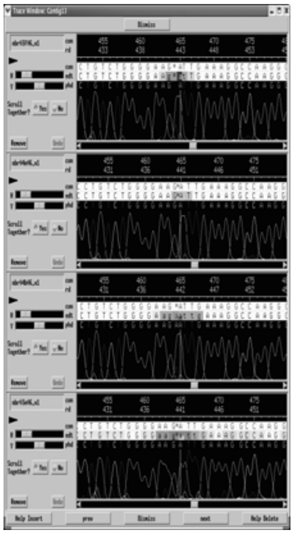
Some of the challenges of variation detection



Heterozygous DIPs



“Dye blob”



Detection saturation

Future of Medical Sequencing

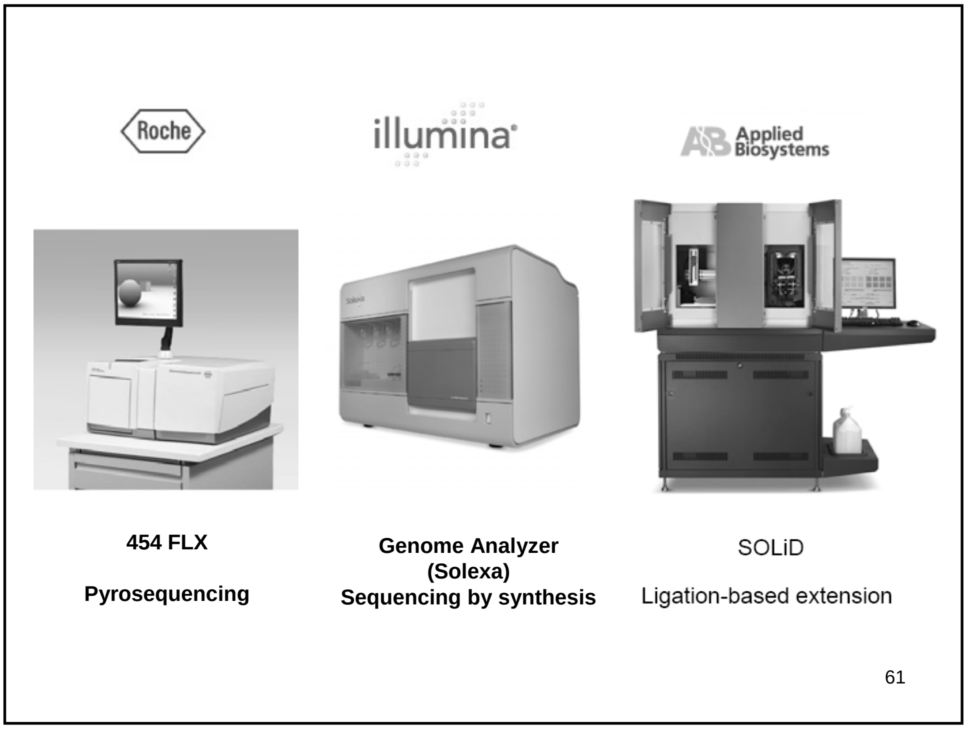
- Many sequencing centers have medical sequencing pipelines in operation
- Next-generation sequencing platforms will radically change this approach

Overview of Topics

- Review of genetic variation discovery
- Database of SNPs, dbSNP
- Other types of genetic variation
- Medical sequencing
- Next-generation sequencing and SNPs
- Targeted Genomic Selection

Next-gen Sequencing

- Introduced by Dr. Margulies in an earlier CTGA lecture
- How these can be used for variation detection and genotyping
- Techniques for targeted genomic capture in combination with next-gen sequencing
- Large scale efforts for greatly expanding the list of known variants in the genome



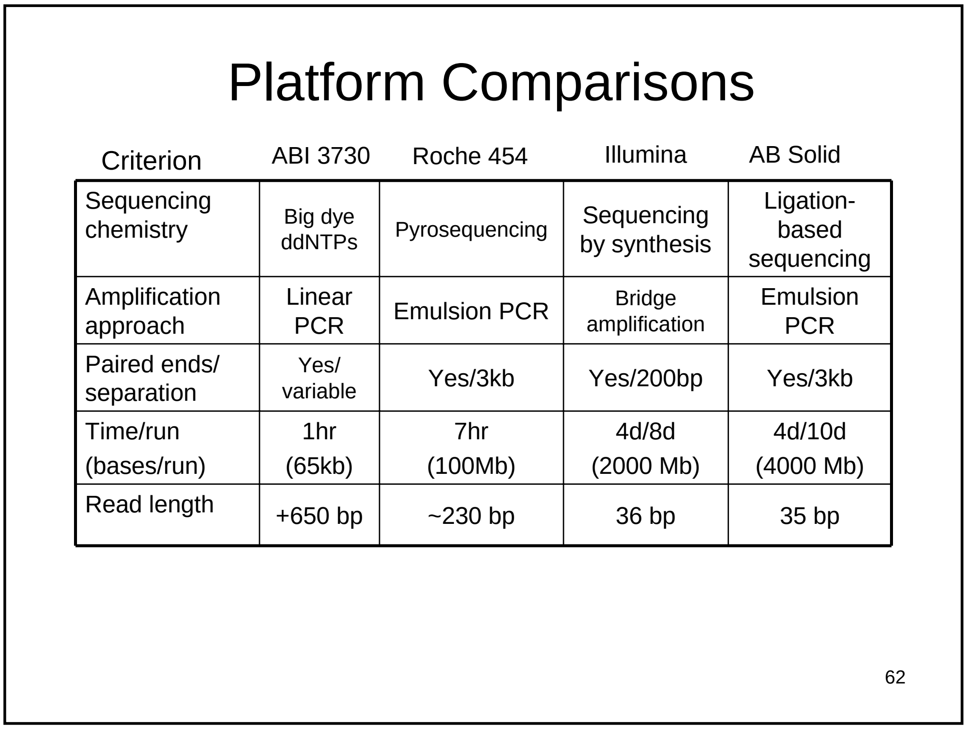
454 FLX

Pyrosequencing

Genome Analyzer (Solexa)
Sequencing by synthesis

SOLiD

Ligation-based extension

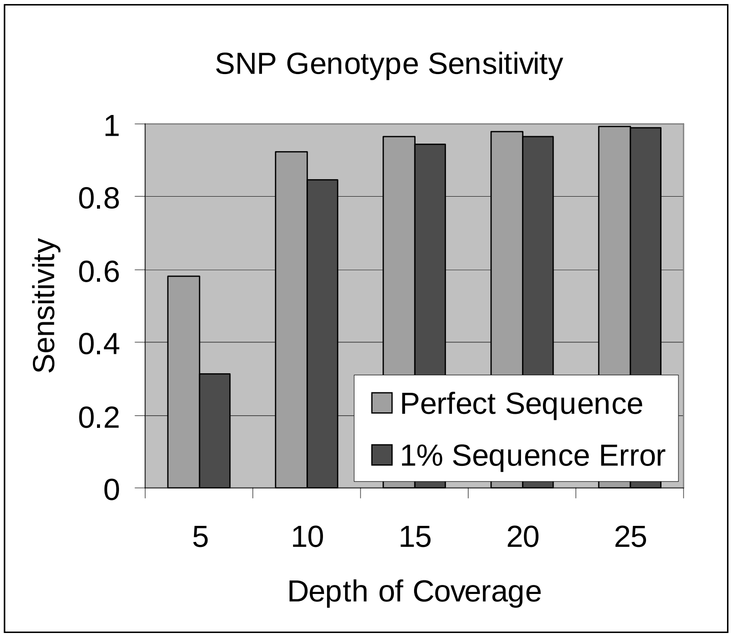


Platform Comparisons

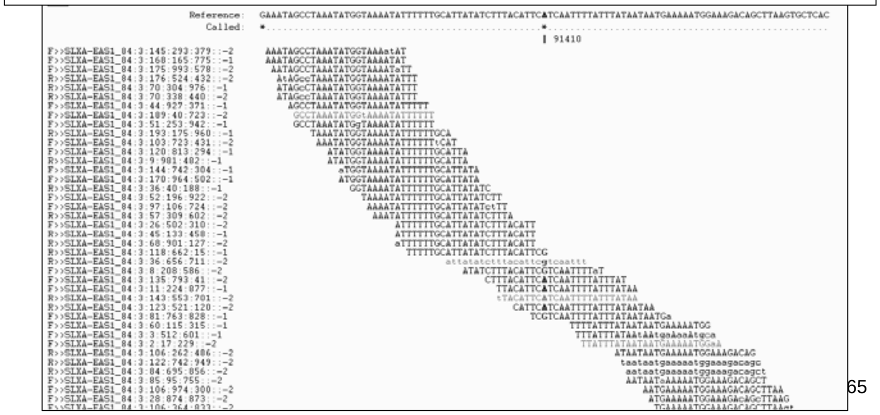
Criterion	ABI 3730	Roche 454	Illumina	AB Solid
Sequencing chemistry	Big dye ddNTPs	Pyrosequencing	Sequencing by synthesis	Ligation-based sequencing
Amplification approach	Linear PCR	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired ends/ separation	Yes/ variable	Yes/3kb	Yes/200bp	Yes/3kb
Time/run (bases/run)	1hr (65kb)	7hr (100Mb)	4d/8d (2000 Mb)	4d/10d (4000 Mb)
Read length	+650 bp	~230 bp	36 bp	35 bp

Next-Gen Sequencing to Detect SNPs from Diploid DNA

- 454-FLX, Solexa and SOLiD generate sequence from clonal substrates
- If one would like to know both alleles at each base, sequence coverage must be high, e.g. over 10X
- To sequence an individual’s diploid genome, therefore, would require at least 30Gb of sequence
  - 300 454-FLX runs (100 machine-days)
  - 15 Solexa runs (120 machine-days)
  - 8 SOLiD runs (80 machine-days)



Example of short read sequence alignment



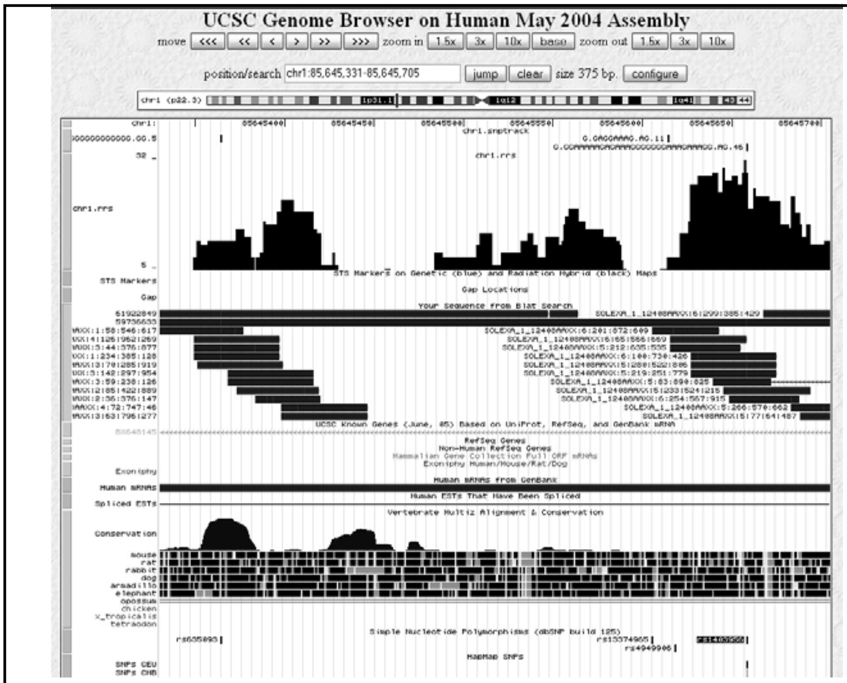
```
                         Reference: GAAATAGCCTAAATATGGTAAAATATTTTTTGCATTATATCTTTACATTCATCAATTTTATTTATAATAATGAAAAATGGAAAGACAGCTTAAGTGCTCAC
                            Called: *.................................................*..................................................
                                                                                      | 91410
F>>SLXA-EAS1_84:3:145:293:379::-2    AAATAGCCTAAATATGGTAAAatAT
F>>SLXA-EAS1_84:3:168:165:775::-1    AAATAGCCTAAATATGGTAAAATAT
F>>SLXA-EAS1_84:3:175:993:578::-2     AATAGCCTAAATATGGTAAAATaTT
R>>SLXA-EAS1_84:3:176:524:432::-2      AtAGccTAAATATGGTAAAATATTT
R>>SLXA-EAS1_84:3:70:304:976::-1       ATAGcCTAAATATGGTAAAATATTT
R>>SLXA-EAS1_84:3:70:338:440::-2       ATAGccTAAATATGGTAAAATATTT
F>>SLXA-EAS1_84:3:44:927:371::-1         AGCCTAAATATGGTAAAATATTTTTT
F>>SLXA-EAS1_84:3:189:40:723::-2          GCCTAAATATGGtAAAATATTTTTT
F>>SLXA-EAS1_84:3:51:253:942::-1          GCCTAAATATGgTAAAATATTTTTT
R>>SLXA-EAS1_84:3:193:175:960::-1            TAAATATGGTAAAATATTTTTTGCA
F>>SLXA-EAS1_84:3:103:723:431::-2             AAATATGGTAAAATATTTTTTtCAT
F>>SLXA-EAS1_84:3:120:813:294::-1               ATATGGTAAAATATTTTTTGCATTA
R>>SLXA-EAS1_84:3:9:901:402::-1                 ATATGGTAAAATATTTTTTGCATTA
F>>SLXA-EAS1_84:3:144:742:304::-1                 aTGGTAAAATATTTTTTGCATTATA
F>>SLXA-EAS1_84:3:170:964:502::-1                 ATGGTAAAATATTTTTTGCATTATA
R>>SLXA-EAS1_84:3:36:40:188::-1                     GGTAAAATATTTTTTGCATTATATC
F>>SLXA-EAS1_84:3:52:196:922::-2                      TAAAATATTTTTTGCATTATATCTT
F>>SLXA-EAS1_84:3:97:106:724::-2                       AAAATATTTTTTGCATTATATctTT
R>>SLXA-EAS1_84:3:57:309:602::-2                        AAATATTTTTTGCATTATATCTTTA
F>>SLXA-EAS1_84:3:26:502:310::-2                            ATTTTTTGCATTATATCTTTACATT
R>>SLXA-EAS1_84:3:45:133:458::-1                            ATTTTTTGCATTATATCTTTACATT
R>>SLXA-EAS1_84:3:68:901:127::-2                            aTTTTTTGCATTATATCTTTACATT
R>>SLXA-EAS1_84:3:118:662:15::-1                              TTTTTGCATTATATCTTTACATTCG
R>>SLXA-EAS1_84:3:36:656:711::-2                                     attatatctttacattcgtcaattt
F>>SLXA-EAS1_84:3:8:208:586::-2                                         ATATCTTTACATTCGTCAATTTTaT
F>>SLXA-EAS1_84:3:135:793:41::-2                                            CTTTACATTCATCAATTTTATTTAT
F>>SLXA-EAS1_84:3:11:224:877::-1                                              TTACATTCATCAATTTTATTTATAA
R>>SLXA-EAS1_84:3:143:553:701::-2                                             tTACATTCATCAATTTTATTTATAA
R>>SLXA-EAS1_84:3:123:521:120::-2                                                CATTCATCAATTTTATTTATAATAA
F>>SLXA-EAS1_84:3:81:763:828::-1                                                    TCGTCAATTTTATTTATAATAATGa
F>>SLXA-EAS1_84:3:60:115:315::-1                                                           TTTTATTTATAATAATGAAAAATGG
F>>SLXA-EAS1_84:3:3:512:601::-1                                                             TTTATTTATAAtAAtgaAaaAtgca
F>>SLXA-EAS1_84:3:2:17:229::-2                                                               TTATTTATAATAATGAAAAATGGaA
R>>SLXA-EAS1_84:3:106:262:486::-2                                                                  ATAATAATGAAAAATGGAAAGACAG
R>>SLXA-EAS1_84:3:122:742:949::-2                                                                   taataatgaaaaatggaaagacagc
R>>SLXA-EAS1_84:3:84:695:856::-2                                                                     aataatgaaaaatggaaagacagct
F>>SLXA-EAS1_84:3:85:95:755::-2                                                                      AATAATaAAAAATGGAAAGACAGCT
F>>SLXA-EAS1_84:3:106:974:300::-2                                                                       AATGAAAAATGGAAAGACAGCTTAA
F>>SLXA-EAS1_84:3:28:874:873::-2                                                                         ATGAAAAATGGAAAGAcAGcTTAAG
F>>SLXA-EAS1_84:3:106:364:833::-2                                                                         TGAAAAATGGAAAGACAGCTTAAgt
```

65

SNP/Genotype Calling

- Alleles at each base with aligned data called using a Bayesian based method
  - ten possible genotypes, four homozygous and 6 heterozygous
  - Non-reference genotype prior probability is 0.001, sequencing error rate is 1.7%
  - Score is the difference between the log-odds of the most probable genotype and the second most probable genotype

66


UCSC Genome Browser on Human May 2004 Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr1:85,645,331-85,645,705 jump clear size 375 bp. configure
chr1 (p22.3) 1p31.1 1q12 1q41 43 44
chr1: 85645400 85645450 85645500 85645550 85645600 85645650 85645700
chr1.snptrack
G.GAGGAAAG.AG.11
G.GGAAAAAGAGAAAGGGGGGGAAAGAAAGG.AG.46
52 _
chr1.rrs
5 _
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Gap Locations
Gap
Your Sequence from Blat Search
61922849
59736633
SOLEXA_1_12408AAXX:6:299:385:429
SOLEXA_1_12408AAXX:6:201:872:609
SOLEXA_1_12408AAXX:6:65:566:669
SOLEXA_1_12408AAXX:5:212:635:535
SOLEXA_1_12408AAXX:6:100:730:426
SOLEXA_1_12408AAXX:5:280:522:806
SOLEXA_1_12408AAXX:5:219:251:779
SOLEXA_1_12408AAXX:5:83:890:825
SOLEXA_1_12408AAXX:5:233:524:215
SOLEXA_1_12408AAXX:6:254:567:915
SOLEXA_1_12408AAXX:5:266:570:662
SOLEXA_1_12408AAXX:5:77:641:487
AXX:1:58:546:617
XX:4:126:962:269
AXX:5:44:376:877
XX:1:234:385:128
AXX:3:70:285:919
XX:5:142:297:954
AXX:3:59:238:126
AXX:2:85:422:889
AXX:2:36:376:147
AAXX:4:72:747:46
AXX:3:63:796:277
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
BX648145
RefSeq Genes
Non-Human RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
Exoniphy Human/Mouse/Rat/Dog
Exoniphy
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & Conservation
Conservation
mouse
rat
rabbit
dog
armadillo
elephant
opossum
chicken
x_tropicalis
tetraodon
Simple Nucleotide Polymorphisms (dbSNP build 125)
rs635093
rs13374965
rs4949906
HapMap SNPs
SNPs CEU
SNPs CHB

67

Overview of Topics

- Review of genetic variation discovery
- Database of SNPs, dbSNP
- Other types of genetic variation
- Medical sequencing
- Next-generation sequencing and SNPs
- Targeted Genomic Selection

68

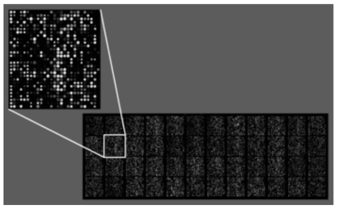
Targeted Genomic Selection

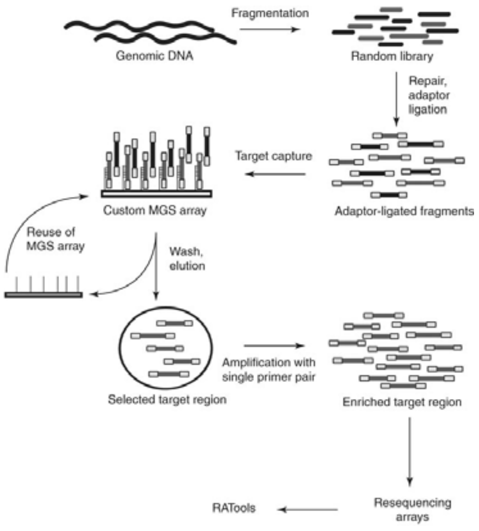
- Multiplex PCR
  - – Expensive to cover large regions
- Reduced representation using restriction enzymes
  - – Inexpensive, but cannot be targeted
- Long Range PCR
  - – Difficult to design, suffers from allelic dropout

- Hybridization capture
- Molecular Inversion Probe capture

Microarray Direct Capture

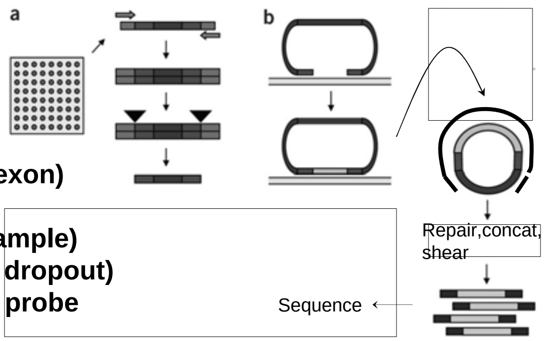
- **385k features / chip (6MB seq.)**
- **24-30k exons / chip**
- **55-85% specificty (seq in ROI)**
- **12-50% total ROI seq coverage***
- **Exon coverage 40-78% (22-60% dropout)**
- **Non-uniform seq. depth**
- **20 ug DNA input**





Hodges etal. *Nature Genetics* **39**, 1522-1527 (2007)
Okou etal *Nature Methods* **4**, 907-909 (2007)

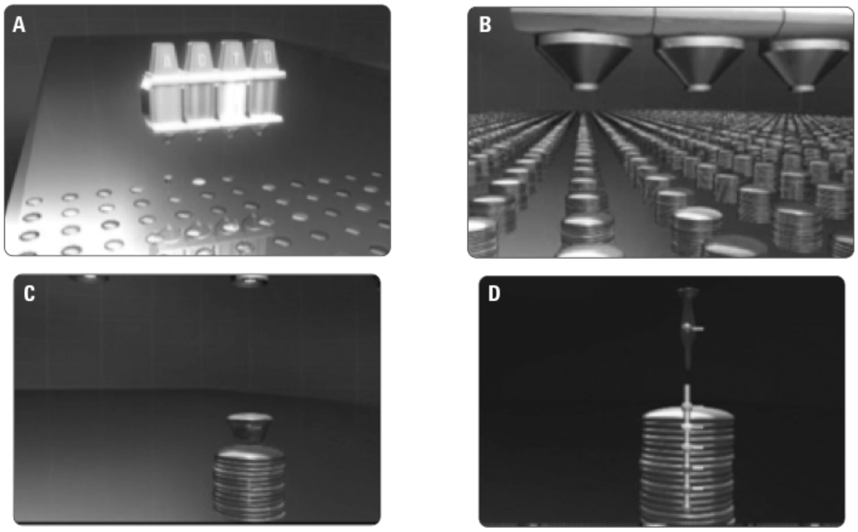
Molecular Inversion Probe Capture



- 55,000 capture oligos (1 / exon)
- 6.7 Mb total seq.
- Specifity = 98.6% (small sample)
- Exon coverage = 91% (9% dropout)
- Each exon targeted with 1 probe
- 750 ng - 1.5 ug DNA input
- Highly non-uniform seq. coverage (several logs) - but consistent
- Het calls - 96% sens.

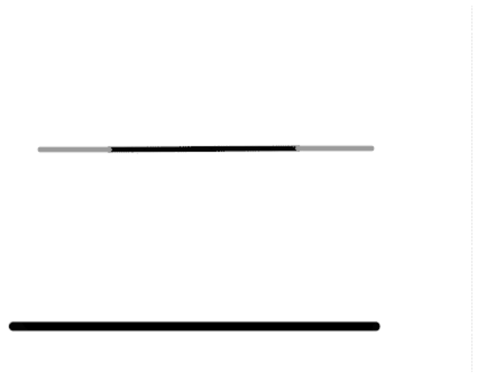
Porreca etal. *Nature Methods* **4**, 931-936 (2007)

High Throughput Synthesis
Of Long Oligo Libraries

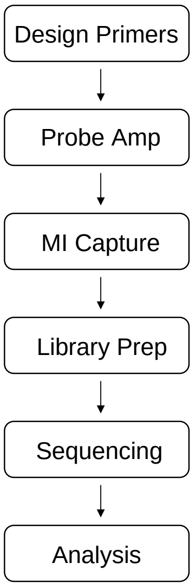


Agilent Technologies

Molecular Inversion Probe Capture



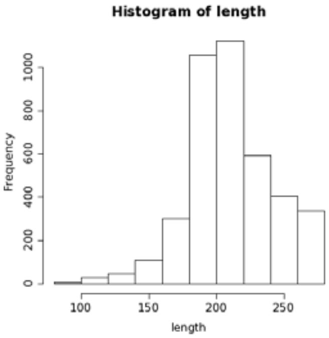
Experimental Design - Workflow

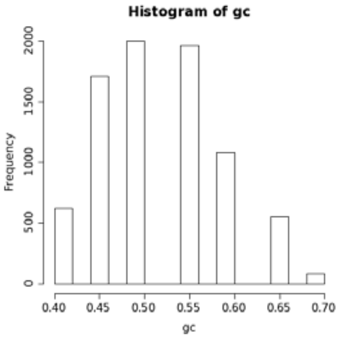


Design Primers	3-4 days
Probe Amp	4 days
MI Capture	3-4 days
Library Prep	1-2 days
Sequencing	3 days
Analysis	

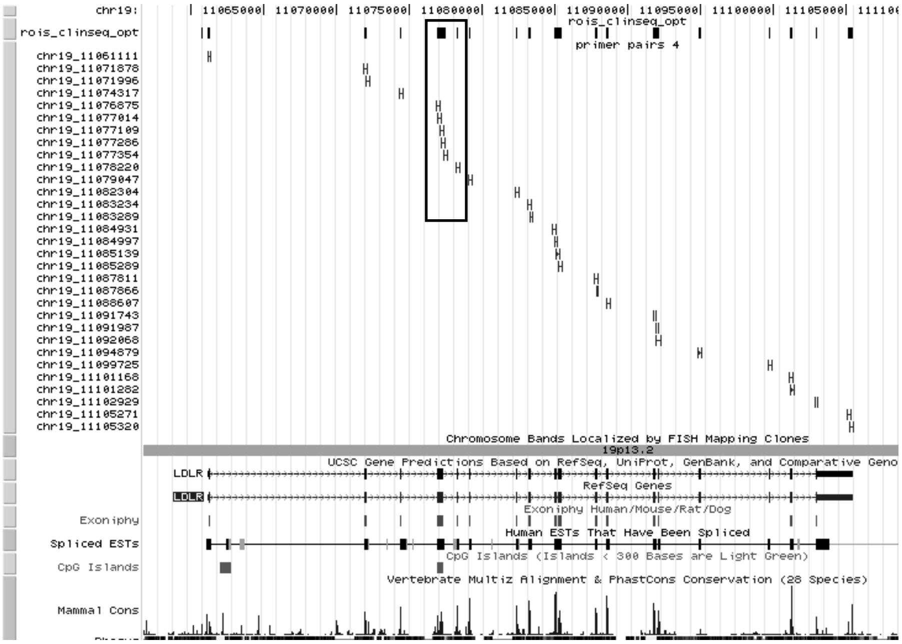
Experimental Design - Primers

- **4000 probes - (492kbp)**
- **Use primer_tile to design - similar to PCR primers**
- **Capture Region Criteria**
  - **length of region: 90-280bp, 200 optimal**
  - **GC% in targeting pairs: 45-65%, then 40-70%**
  - **minimize non-specific targeting pairs using ePCR**
  - **no Nt.AlwI or Nb.BsrDI restriction sites in targeting arms**
  - **no SNPs in targeting arms**

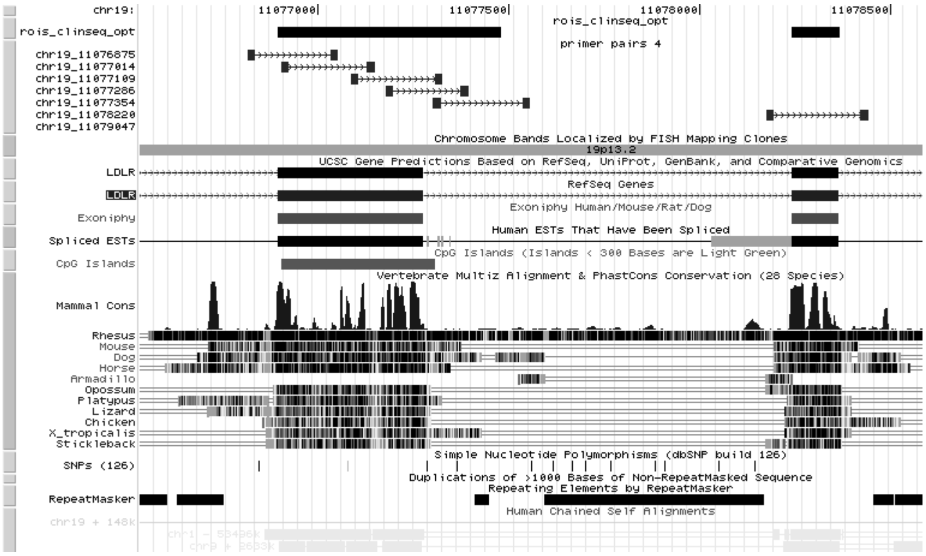




Example Region

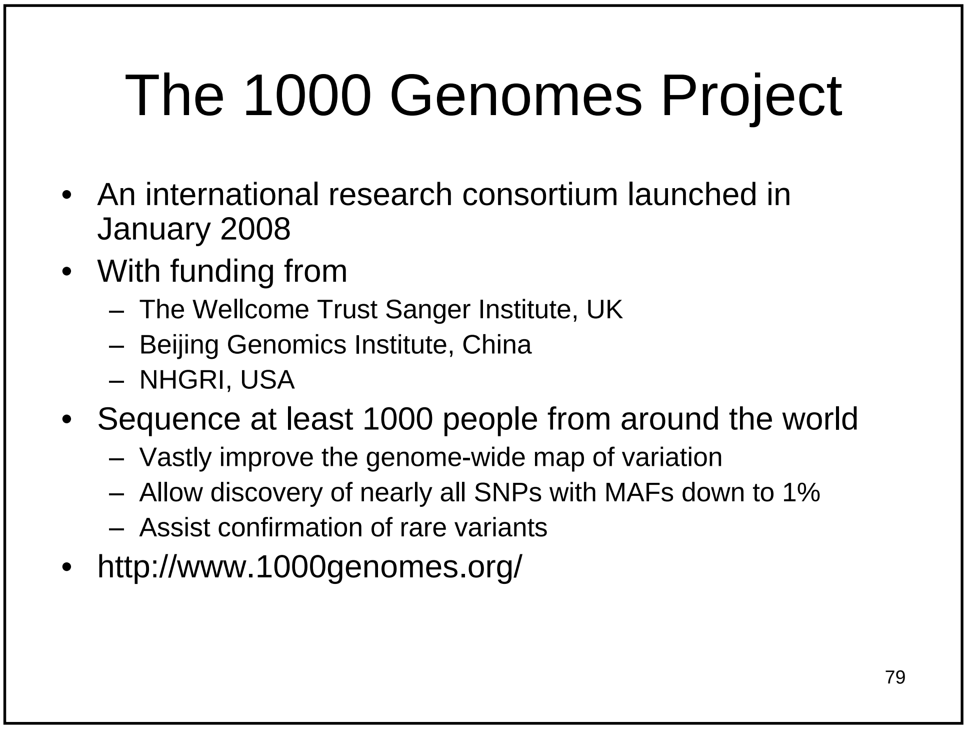


Example Zoomed In


chr19:
11077000|
11077500|
11078000|
11078500|
rois_clinseq_opt
rois_clinseq_opt
primer pairs 4
chr19_11076875
chr19_11077014
chr19_11077109
chr19_11077286
chr19_11077354
chr19_11078220
chr19_11079047
Chromosome Bands Localized by FISH Mapping Clones
19p13.2
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
LDLR
RefSeq Genes
LDLR
Exoniphy Human/Mouse/Rat/Dog
Exoniphy
Human ESTs That Have Been Spliced
Spliced ESTs
CpG Islands (Islands < 300 Bases are Light Green)
CpG Islands
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 126)
SNPs (126)
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Repeating Elements by RepeatMasker
RepeatMasker
Human Chained Self Alignments
chr19 + 148k

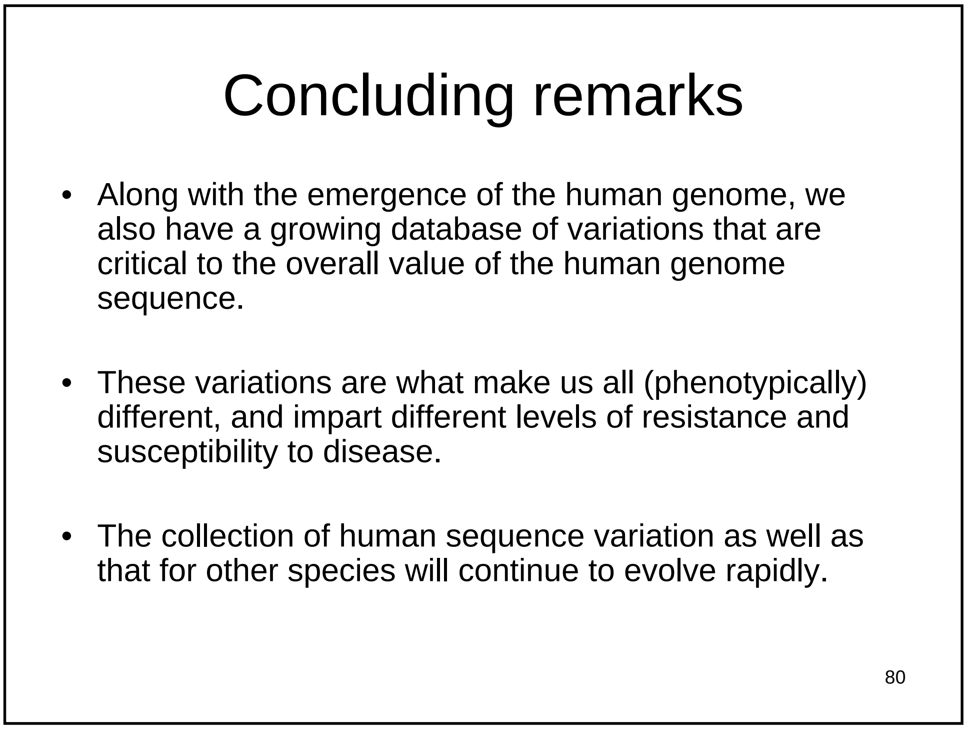
Which technology to use depends on the scale of the project

- PCR with Sanger based sequencing
  - 10s of exons
  - 250 amplicons
- Targeted genomic selection and next-gen sequencing
  - Over 2Mb of sequence
  - Entire exome
  - Part of a chromosome



The 1000 Genomes Project

- An international research consortium launched in January 2008
- With funding from
  - The Wellcome Trust Sanger Institute, UK
  - Beijing Genomics Institute, China
  - NHGRI, USA
- Sequence at least 1000 people from around the world
  - Vastly improve the genome-wide map of variation
  - Allow discovery of nearly all SNPs with MAFs down to 1%
  - Assist confirmation of rare variants
- http://www.1000genomes.org/

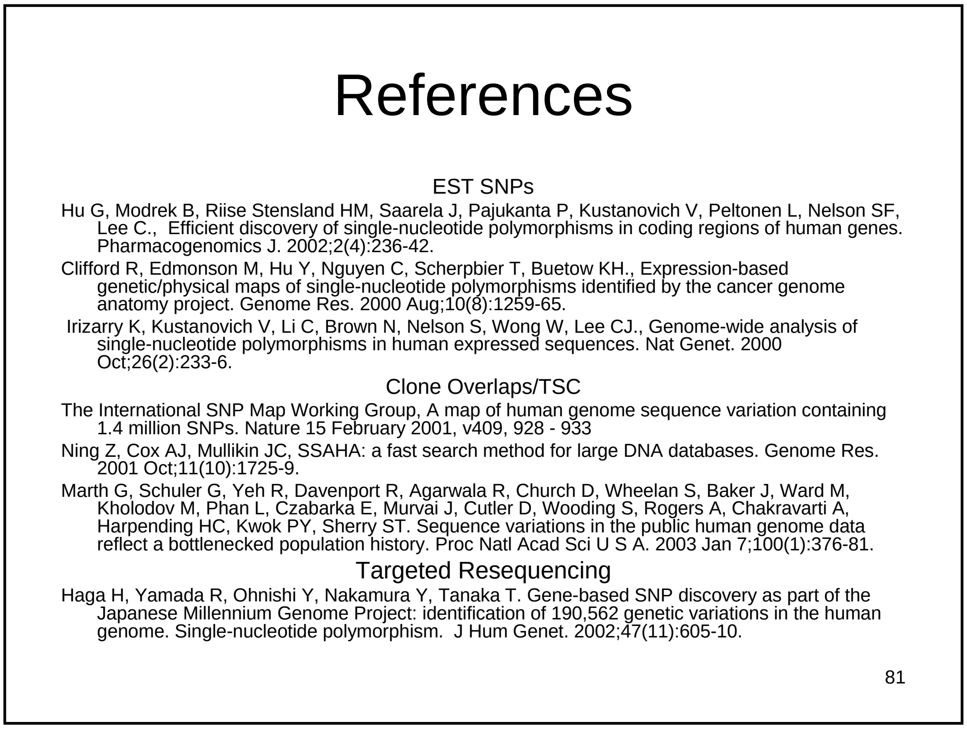


Concluding remarks

- Along with the emergence of the human genome, we also have a growing database of variations that are critical to the overall value of the human genome sequence.

- These variations are what make us all (phenotypically) different, and impart different levels of resistance and susceptibility to disease.

- The collection of human sequence variation as well as that for other species will continue to evolve rapidly.



References

EST SNPs

Hu G, Modrek B, Riise Stensland HM, Saarela J, Pajukanta P, Kustanovich V, Peltonen L, Nelson SF, Lee C., Efficient discovery of single-nucleotide polymorphisms in coding regions of human genes. Pharmacogenomics J. 2002;2(4):236-42.

Clifford R, Edmonson M, Hu Y, Nguyen C, Scherpbier T, Buetow KH., Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. Genome Res. 2000 Aug;10(8):1259-65.

Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W, Lee CJ., Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. Nat Genet. 2000 Oct;26(2):233-6.

Clone Overlaps/TSC

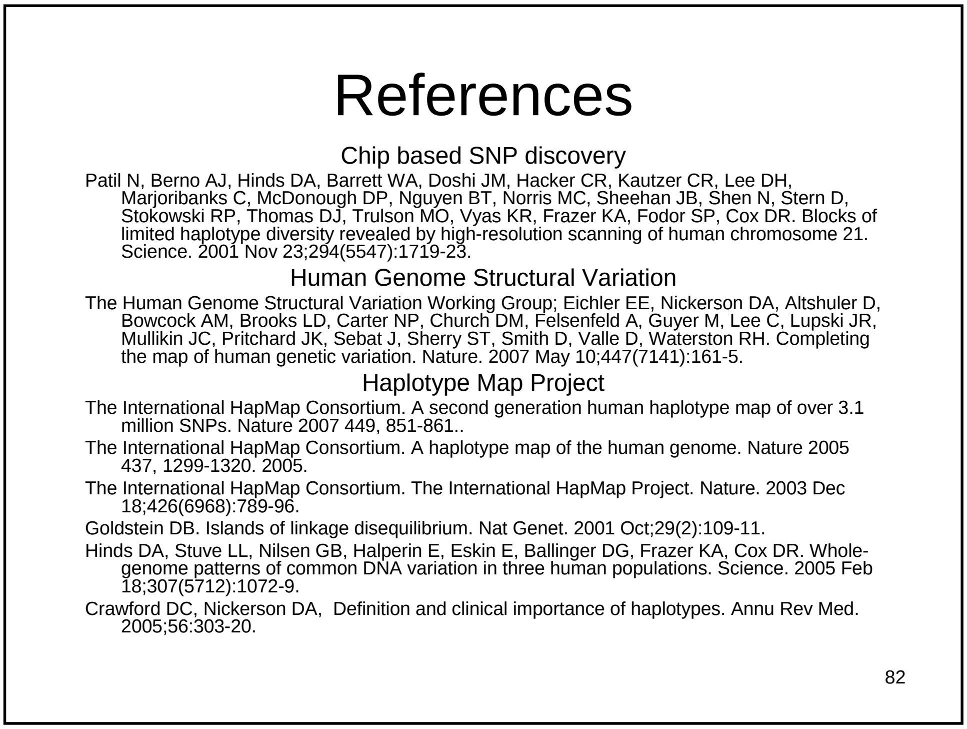
The International SNP Map Working Group, A map of human genome sequence variation containing 1.4 million SNPs. Nature 15 February 2001, v409, 928 - 933

Ning Z, Cox AJ, Mullikin JC, SSAHA: a fast search method for large DNA databases. Genome Res. 2001 Oct;11(10):1725-9.

Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, Church D, Wheelan S, Baker J, Ward M, Kholodov M, Phan L, Czabarka E, Murvai J, Cutler D, Wooding S, Rogers A, Chakravarti A, Harpending HC, Kwok PY, Sherry ST. Sequence variations in the public human genome data reflect a bottlenecked population history. Proc Natl Acad Sci U S A. 2003 Jan 7;100(1):376-81.

Targeted Resequencing

Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. J Hum Genet. 2002;47(11):605-10.



References

Chip based SNP discovery

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science. 2001 Nov 23;294(5547):1719-23.

Human Genome Structural Variation

The Human Genome Structural Variation Working Group; Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, Carter NP, Church DM, Felsenfeld A, Guyer M, Lee C, Lupski JR, Mullikin JC, Pritchard JK, Sebat J, Sherry ST, Smith D, Valle D, Waterston RH. Completing the map of human genetic variation. Nature. 2007 May 10;447(7141):161-5.

Haplotype Map Project

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007 449, 851-861..

The International HapMap Consortium. A haplotype map of the human genome. Nature 2005 437, 1299-1320. 2005.

The International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789-96.

Goldstein DB. Islands of linkage disequilibrium. Nat Genet. 2001 Oct;29(2):109-11.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. Whole-genome patterns of common DNA variation in three human populations. Science. 2005 Feb 18;307(5712):1072-9.

Crawford DC, Nickerson DA, Definition and clinical importance of haplotypes. Annu Rev Med. 2005;56:303-20.

WEB pages

http://droog.mbt.washington.edu/PolyPhred.html
http://www.ncbi.nlm.nih.gov/SNP/index.html : dbSNP home page
http://www.ensembl.org : Ensembl home page
http://www.ucl.ac.uk/~ucbhdjm/courses/b242/2+Gene/2+Gene.html
http://www.hapmap.org/: Haplotype Map Project home page
http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap
http://www.broad.mit.edu/personal/jcbarret/haploview/
http://genome.perlegen.com/browser/index_v2.html: Perlegen’s HapMap
http://www.genome.gov/25521748 : HGSV