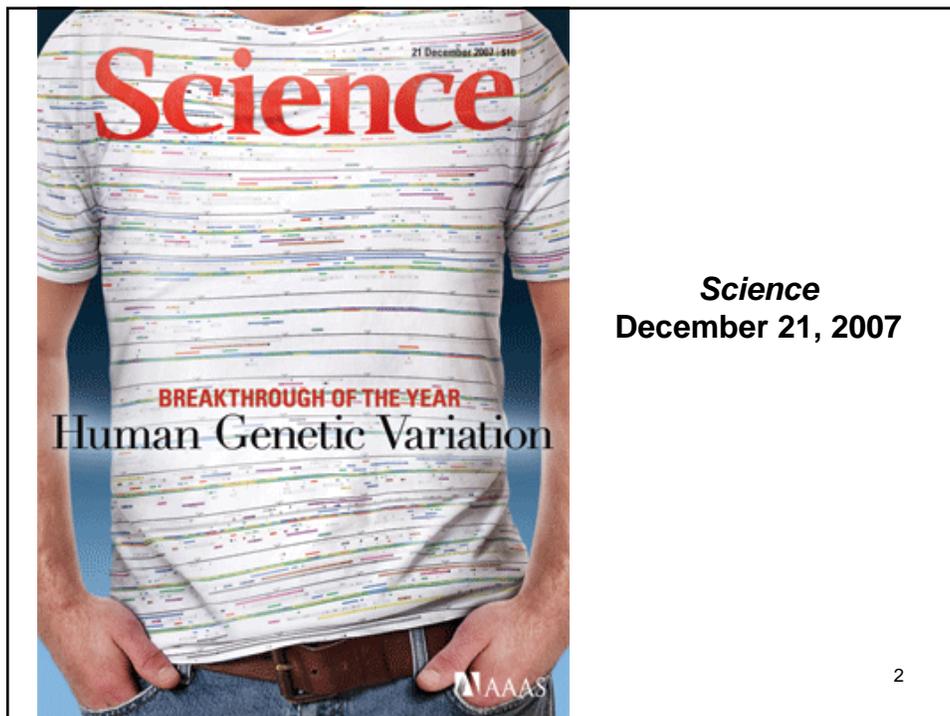


# Studying Genetic Variation II: Computational Techniques

**Jim Mullikin, PhD**  
**Genome Technology Branch**  
**NHGRI**



## Some points from other lectures

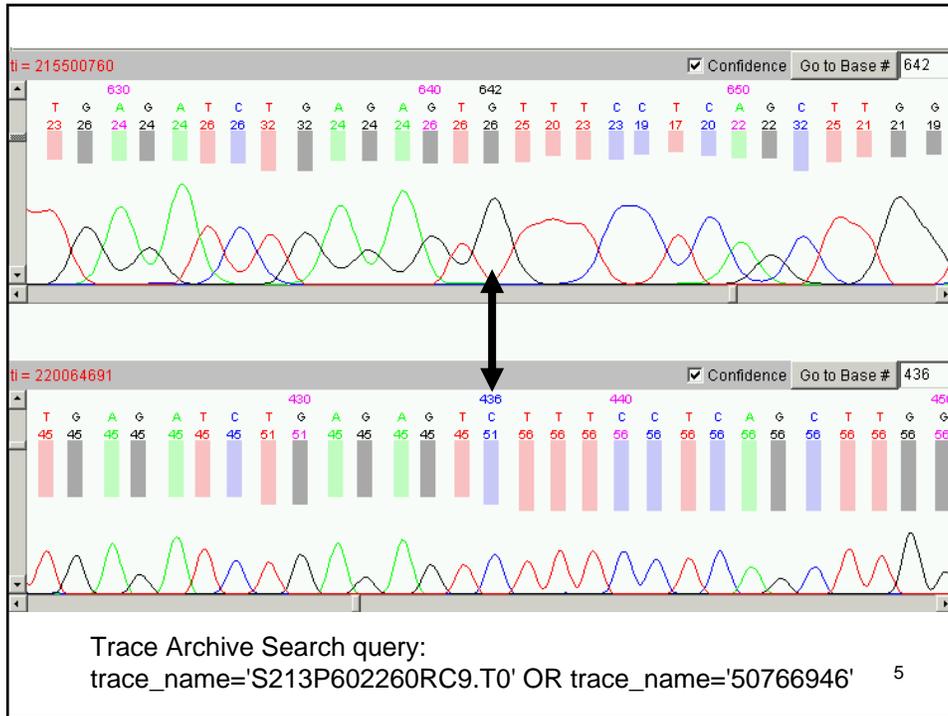
- Population Genetics: Practical Applications  
Lynn Jorde
  - Described patterns of human genetic variation among and within populations, linkage disequilibrium and HapMap and how all this relates to the search for complex disease genes.
- Linkage Analysis and Complex Traits  
Elaine Ostrander
  - Linkage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

3

## Genetic Variation Discovery

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

4



## 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%

7

## 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.

8

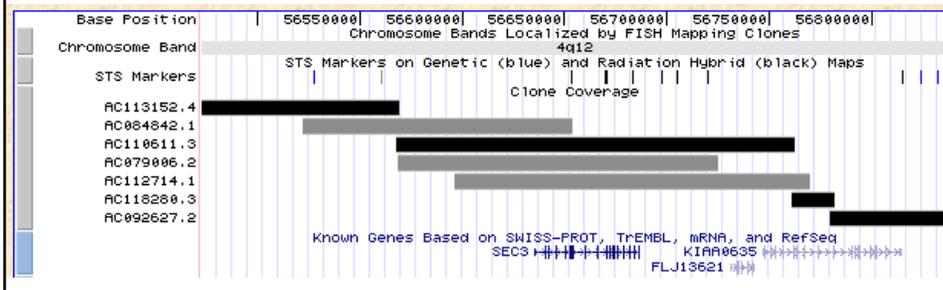
## Review of Genetic Variation Discovery Efforts

- 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.

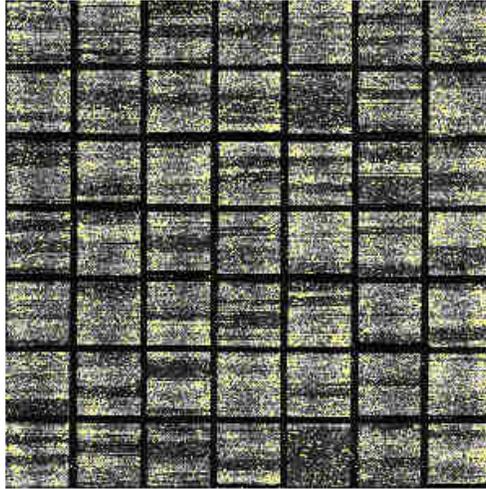
11

## 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.

12

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/>

13

## 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.

14

## 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.

15

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	$\geq 0.05$	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	$\geq 0.05$	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	$\geq 0.05$	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

**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.

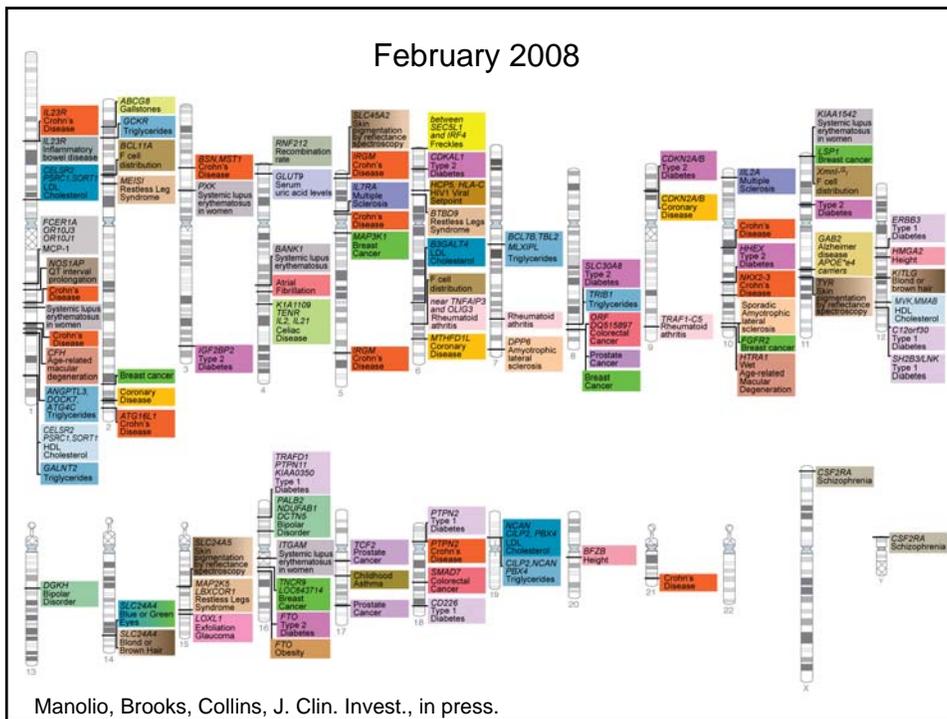
Platform*	CHB+JPT	
	$r^2 \geq 0.8$ (%)	Mean maximum $r^2$
Affymetrix GeneChip 500K	67	0.80
Affymetrix SNP Array 6.0	81	0.89
Illumina HumanHap300	63	0.78
Illumina HumanHap550	83	0.89
Illumina HumanHap650Y	84	0.90
Perlegen 600K	84	0.90

NATURE Vol 449, 18 October 2007 17

## 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](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



# 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 <a href="#">Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33</a>	Breast cancer	249 cases, 299 controls	1,193 cases, 1,166 controls	Affymetrix [391,467]
Krov March 11, 2008 Mol Psychiatry <a href="#">A genome-wide association study in 574 schizophrenia trios using DNA pooling</a>	Schizophrenia	605 controls 574 cases, 1148 parents of cases	NR	Affymetrix (~550,000) (pooled)
Doring March 09, 2008 Nat Genet <a href="#">SLC2A9 influences uric acid concentrations with pronounced sex-specific effects</a>	Uric acid	1,644 individuals	9,947 individuals	Affymetrix [335,152]
Vitart March 09, 2008 Nat Genet <a href="#">SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout</a>	Serum urate	704 individuals	706 individuals	Illumina [308,140]
Liu March 05, 2008 Hum Mol Genet <a href="#">Genome-wide association scans identified CTNRF1 as a novel gene for obesity</a>	Obesity	1,000 individuals	896 obese individuals, 2,916 lean individuals	Affymetrix [379,319]
Sklar March 04, 2008 Mol Psychiatry <a href="#">Whole-genome association study of bipolar disorder</a>	Bipolar disorder	1,461 cases, 2,008 controls	409 trios, 365 cases, 351 controls	Affymetrix [372,193]

And 130 more entries...

20

## How to Interpret a Genome-wide Association Study

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

IN THE PAST 2 YEARS, THERE HAS BEEN a dramatic increase in genomic discoveries involving 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.<sup>1</sup> Nearly 12 million unique human SNPs have been assigned a reference SNP (rs) number in the National Center for Biotechnology Information's dbSNP database<sup>2</sup> and

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

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

21

## dbGaP

- [http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap\\_tmpl/about.html](http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap_tmpl/about.html)
- The **database of Genotype and Phenotype (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](http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap_tmpl/dbGaP_HowTo.pdf)

22

## 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

23

## 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.

24

NCBI Single Nucleotide Polymorphism

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNPs

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for [ ] Go

**BUILD 128**  
Have a question about dbSNP? Try searching the SNP FAQ Archive!  
[ ] Go

**GENERAL**  
Contact Us  
Site Map **dbSNP Homepage**  
Announcements  
dbSNP Summary  
FTP Download  
SNP SUBMISSION  
DOCUMENTATION  
SEARCH  
RELATED SITES

**dbSNP Search Options**  
Entrez SNP ID Numbers Submission Info Batch Locus Info Between Markers

**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)  
[ ] Reference cluster ID(rs#) [v]  
[ Search ] [ Reset ]

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

Single Nucleotide Polymorphism

All Databases PubMed Nucleotide Protein Genom

Search SNP for [ ] Go Clear

Limits Preview/Index History Clipboard Details

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

**SNP**

**NCBI**  
dbSNP BUILD 128

**Entrez SNP**  
Search SNP  
Search Mouse SNP  
Common Query Filters  
Entrez Batch Query  
SNP Link Datamodel

**My NCBI**  
My NCBI help

**Entrez SNP Help**  
Searchable FAQ  
Search Fields  
Programming Utilities  
Batch Report  
Legend  
Examples  
dbSNP Home Page  
Overview

**Entrez Help**  
General help

<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:	Updated search terms
January 5, 2005	
August 14, 2002	Add contig position tag [C:TPOS]

Below are search examples and available search fields.

**Search using wild-card(\*), ranging(), AND, OR, and NOT operators:**

Example	Description
<a href="#">BRC*[Gene Name]</a>	Search SNPs on all genes with names starting with the letter 'BRC' (ie. BRCA1 and BRCA2)
<a href="#">1-5[HEI]</a>	Search SNPs with heterozygosity between 1 and 5 percent
<a href="#">coding nonsynonymous[FUNCTION] [CHR]</a>	Search SNPs with function class 'coding nonsynonymous' located on chromosome 1
<a href="#">1[CHR] OR 2[CHR]</a>	Search all SNPs on chromosome 1 or 2
<a href="#">1[CHR] OR 2[CHR] NOT unknown[METHOD]</a>	Search all SNPs on chromosome 1 or 2 detected by all methods except 'unknown'.
<a href="#">1[WEIGHT] AND 1[CHR] OR 2[CHR] NOT unknown[METHOD] OR computed[METHOD]</a>	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 used for querying SNP (e.g. [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], [VAR]	JUPAC	Observed allele(s) Example: <a href="#">N[ALLELE]</a>
Chromosome	[CHR]	Textnum	Mapped chromosome number Available values [1-22,W-Z, and Un (unknown)] Example: <a href="#">2[CHR]</a> or <a href="#">X[CHR]</a>
Base Position	[CHRPOS][BPOS]	Integer*	Mapped chromosome position; use in conjunction with chromosome field [CHR] Example: <a href="#">7[CHR] AND 88556198-88580393[CHRPOS]</a>
Create Build ID	[CREATE_BUILD][CBID]	Integer*	SNP create build ID Example: <a href="#">103[CBID]</a>
Publication Date	[CREATEDATE][CDAT],[PDAT],[PUBDATE]	Date*	SNP create publication date Use the format YYYYMMDD, month and day are optional. Example: <a href="#">2005-07-13[CDATE]</a>
Function Class	[FXN_CLASS], [FUNC]	Text	Function Class: locus region coding nonsynonymous coding synonymous ..... intron mRNA utr reference .....

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

27

# dbSNP record for rs1045012

**Reference SNP(refSNP) Cluster Report: rs1045012**

refSNP ID: rs1045012	Allele	Links - Linkout
Organism: human ( <i>Homo sapiens</i> )	Variation Class: SNP	
Molecule Type: Genomic	single nucleotide polymorphism	
Created/Updated in build: 86/128	Alleles: C/G	
Map to Genome Build: <a href="#">36.2</a>	Ancestral Allele: C	

SNP Details are organized in the following sections:

[Submission](#) [Fasta](#) [Resource](#) [GeneView](#) [Map](#) [Diversity](#) [Validation](#)

**Submitter records for this RefSNP Cluster**

The submission [ss44782239](#) 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
<a href="#">ss1514795</a>	LEE e151902		revT	C/G	caacaacccatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	09/13/00	10/10/03	86
<a href="#">ss2423651</a>	HGBASE SNP00010888		revT	C/G	accatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	11/07/00	10/10/03	89
<a href="#">ss2733260</a>	TSC-CSHL TSC0484041		fvdB	C/G	ctctgacaccttggccatttggccacct	ttttcatagatatgacctcatgttggttg	01/02/01	10/10/03	92
<a href="#">ss4391917</a>	LEE e151903		revT	C/G	caacaacccatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	04/25/02	10/10/03	106
<a href="#">ss4407741</a>	LEE e151902		revT	C/G	caacaacccatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	04/26/02	10/10/03	106
<a href="#">ss5815409</a>	SC_JCM NT_007933.10_24217856		revT	C/G	caacaacccatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	01/10/03	10/10/03	111
<a href="#">ss14546249</a>	WUGSC_SSAHASNP chr7.NT_007933.13_24217938		revT	C/G	caacaacccatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	11/05/03	11/02/05	120
<a href="#">ss16262424</a>	CGAP-GAU 1452089		revT	C/G	caacaacccatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	11/18/03	11/22/03	121
<a href="#">ss23476704</a>	PERLEGEN af0546573	<input checked="" type="checkbox"/>	revT	C/G	caacaacccatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	08/10/04	09/13/04	123
<a href="#">ss44782239</a>	AB bc CVS303492	<input checked="" type="checkbox"/>	revT	C/G	caacaacccatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	07/19/05	11/03/06	126
<a href="#">ss48417634</a>	APPLERA_G bc CVS303492	<input checked="" type="checkbox"/>	fvdB	C/G	ctctgacaccttggccatttggccacct	ttttcatagatatgacctcatgttggttg	09/28/05	11/03/06	126
<a href="#">ss69023396</a>	PERLEGEN PGP00546573	<input checked="" type="checkbox"/>	revT	C/G	caacaacccatgaggtgcatactatgaaa	agcgtgccaattggccaaggtgcaacag	01/30/07	08/14/07	127

28

**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
CAACCTCTG AGTTTTTGGT AACATGAGCC AACACAATCC CCTTAAAAAT GAACCACTT
TSAATCCGGG TTTCAGGGTG AGTGGGAGGA TGCTCCAGAA TGAGTGGCCA TGCCCTGCTT
TCCACCACC CCCCAACCCA CCACTCTCTT TCAGGACGGT GGTOCCAGCC ACCCTGACAT
ACCTGTCAAC TGCCGTTTGT GCTCCTTGGG CTGGTGCACC TTGGTCCATT TGCCACCCTG
S
TTTTCATAGA TATGCACCTC ATGGTTGTTG GGGCAGATGG CAATCTCTGA AGGGGAGATG
GAGGGAGATT GAGGGGCCCT CTCCTGACT GCCCTCTGCC AGGACACACT ACACAGTGCA
CCTAGGCAAC AACACCTCAC CTTTCATGAC TCAGTCTCTC CTCTTCTGCC TTGCAGGGGC
CCCTGAAAT CCTTCAGGCC CTGCTAGGCC ACCCTGTCTT CTCTGGAAAC TGCGTGTCTT
TTACTGGCAG CAATGAACCC TGGGACCTCT CCCACCCCTA TTGCTCTGGC CAACCAGGAA
```

**GeneView**

GeneView via analysis of contig annotation: [ARPC1B](#) actin related protein 2/3 complex, subunit 1B, 41kDa  
 Click to see [\[all\]](#) [\[cSNP\]](#) [\[has frequency\]](#) [\[double hit\]](#) [\[haplotype tagged\]](#) variations associated with this gene.

Group Label	Contig->mRNA	Gene Model (contig mRNA transcript) <a href="#">Color Legend</a>
reference	<a href="#">NT_007933-&gt;NM_005720</a> <a href="#">sv function</a>	
Celera	<a href="#">NW_923574-&gt;NM_005720</a> <a href="#">sv function</a>	
CRA_TCAChr7v2	<a href="#">NT_079595-&gt;NM_005720</a> <a href="#">sv function</a>	

Group label	Contig->mRNA->Protein	Contig position	mRNA orientation	mRNA pos	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
reference	<a href="#">NT_007933-&gt;NM_005720-&gt;NP_005711</a>	24218630	forward	200	nonsynonymous	C	Asn [N]	3	37
					contig reference	G	Lys [K]	3	37
Celera	<a href="#">NW_923574-&gt;NM_005720-&gt;NP_005711</a>	22257590	forward	200	nonsynonymous	C	Asn [N]	3	37
					contig reference	G	Lys [K]	3	37
CRA_TCAChr7v2	<a href="#">NT_079595-&gt;NM_005720-&gt;NP_005711</a>	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	<a href="#">NW_923574.1</a>	22257590	93718553	minus	G	alt_assembly_1	Celera	Celera	<a href="#">view</a>	300
7	<a href="#">NT_079595.2</a>	24245339	98344127	minus	G	alt_assembly_2	CRA_TCAChr7v2	CRA_TCAChr7v2	<a href="#">view</a>	300
7	<a href="#">NT_007933.14</a>	24218630	98822290	minus	G	ref_assembly	reference	reference	<a href="#">view</a>	300

**NCBI Resource Links**

Submitter-Referenced	dbSNP Blast Analysis	UniGene Cluster ID	3D structure mapping
GenBank <a href="#">T74087</a> <a href="#">BM803458</a> <a href="#">Hs.11538</a>	GenBank HTGS Finished: <a href="#">AC004922.2</a> <a href="#">NC_000007.12</a>	<a href="#">489284</a>	<a href="#">NP_005711</a>

**Population Diversity**

ss#	Population	Sample Ascertainment				Source	Genotypes			Alleles	
		Individual Group	Sample (2N)	Founder (N)	C/C		C/G	HWP	C	G	Het. +/-std err
<a href="#">ss23476794</a>	<a href="#">AFD_EUR_PANEL</a>	European	48	24	IG	0.917	0.083	1.000	0.938	0.042	
	<a href="#">AFD_AFR_PANEL</a>	African American	46	23	IG	0.739	0.261	0.479	0.870	0.130	
	<a href="#">AFD_CHN_PANEL</a>	Asian	48	24	IG	0.958	0.042	1.000	0.979	0.021	
<a href="#">ss44782239</a>	<a href="#">AoD_African_American</a>		90		AF				0.880	0.120	

SNP ID	Population	n	IG	C	T	G	A	0.990	0.010
ss48417634	African American	78	IG	0.795	0.205	0.479	0.897	0.103	
ss69023396	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	RefSNP Genotype Summary	Total Individual	C/C	C/G	G/G
ss23476794	71	9	62		rs1045012	371	36	53	281
ss44782239	269	5	37	224					
ss48417634	39	31	8						
ss69023396	269	5	37	227					

Discordant Genotypes:							
Individual 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

31

# Viewing SNPs in Browsers

NCBI                      Ensembl                      UCSC

**UCSC Genome Browser on Human Mar. 2006 Assembly**

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

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

The screenshot displays the UCSC Genome Browser interface for a specific region on chromosome 7. At the top, navigation controls allow for zooming in and out of the genomic view. Below the navigation, the current position and search criteria are shown. The main track area contains several layers of information: STS markers, RefSeq genes, Conservation scores, and a detailed sequence view. The sequence view includes annotations for various features, such as SNPs (rs1045012 and rs2177106), repeats, and self-alignments. The interface is designed to provide a comprehensive view of the genomic data, including sequence, annotations, and conservation scores.

## 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

33

## 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:

```
atttattttattttt reference
attt----attttatt rs10581774
attttattta----ttt rs34505627
```

- Structural Variation
- Copy Number Variation

34

## 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)

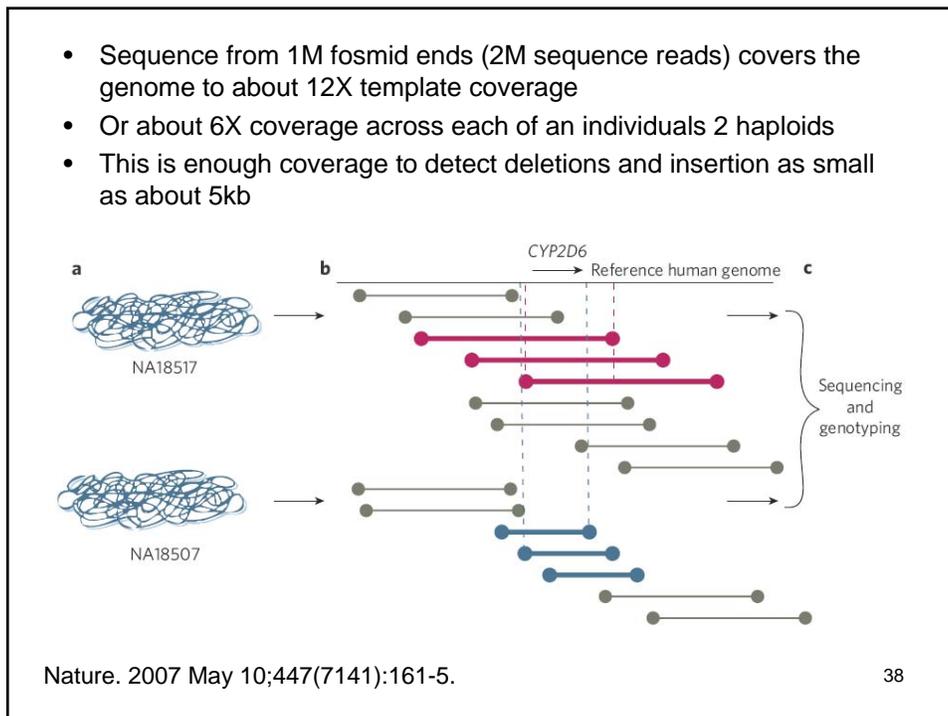
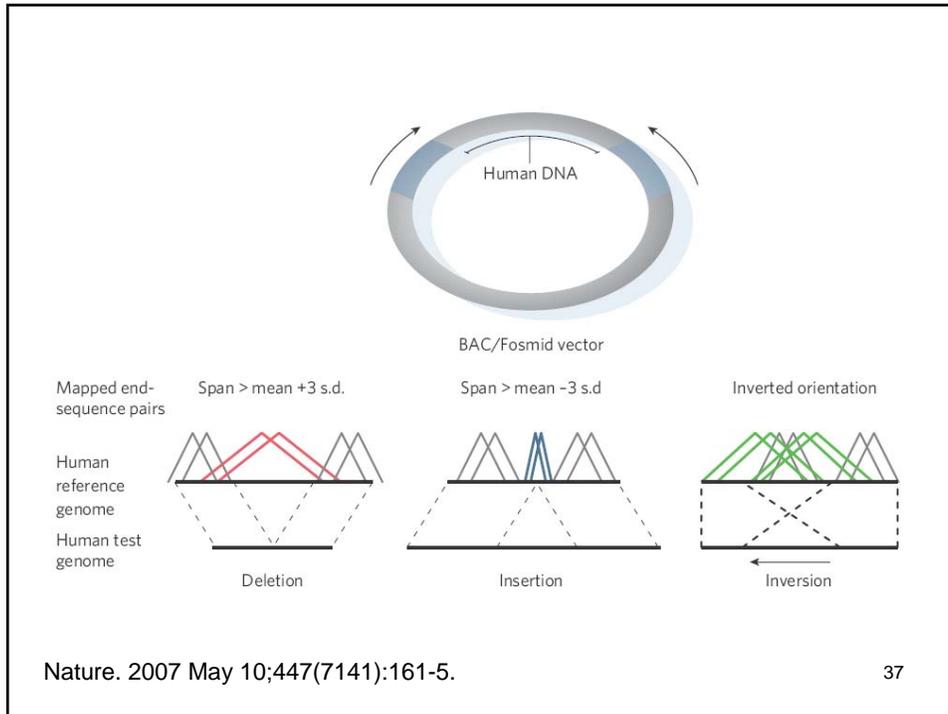
*Genome Res.* 2006 16: 949-961

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



**Table 1 | Common structural polymorphisms and disease**

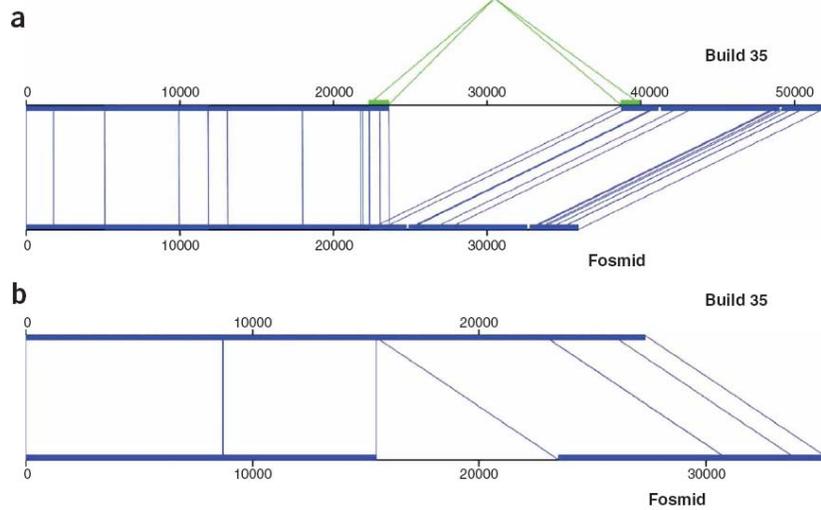
Gene	Type	Locus	Size (kb)	Phenotype	Copy number variation
UGT2B17	Deletion	4q13	150	Variable testosterone levels, risk of prostate cancer	0-2
DEFB4	VNTR	8p23.1	20	Colonic Crohn's disease	2-10
FCGR3	Deletion	1q23.3	>5	Glomerulonephritis, systemic lupus erythematosus	0-14
OPN1LW/OPN1MW	VNTR	Xq28	13-15	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 susceptibility	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.

39

### Sequence level identification of deletion and insertion events



Nature Genetics 37, 727 - 732 (2005)

40

## 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

41

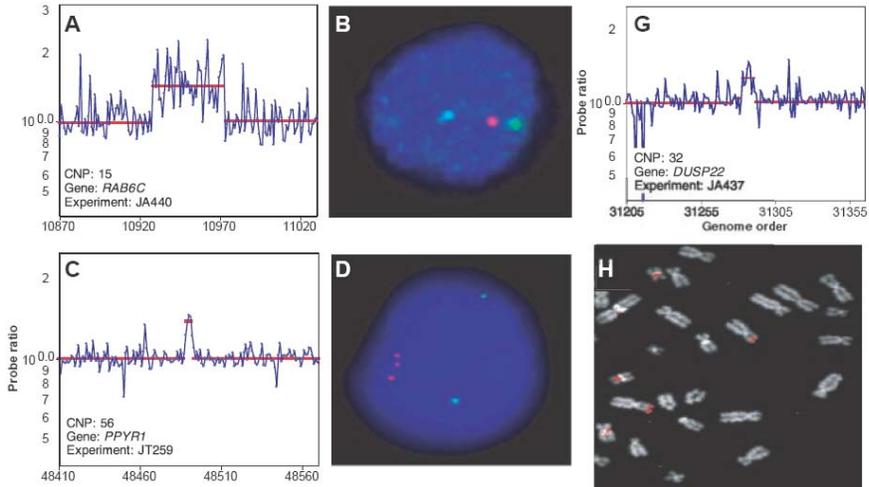
<http://www.genome.gov/25521748>

## 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

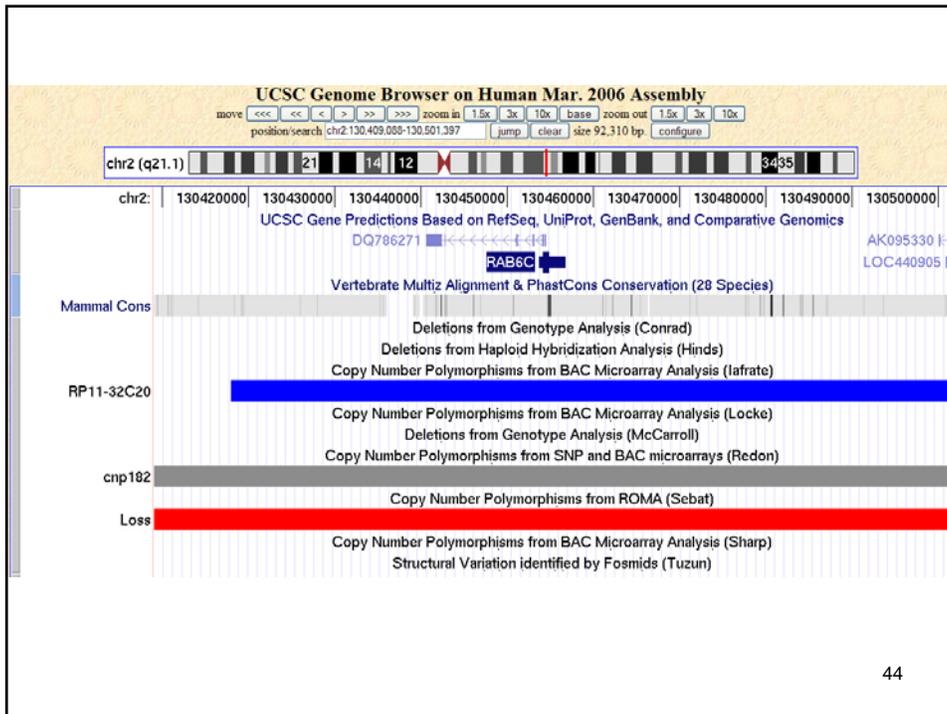
42

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



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

43



44

## 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>

47

- 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; Aagaens Syndrome)	Laura Bull	<a href="#">214900</a>	<a href="#">WUGSC</a>	Assigned
Mendelian	Joubert Syndrome (JBS1)	Joseph Gleeson	<a href="#">213300</a>	<a href="#">RI-MIT</a>	Assigned
Mendelian	Dominant Restrictive Cardiomyopathy	Margart Wallace	<a href="#">609578</a>	<a href="#">NISC</a>	Assigned
Mendelian	Thoracic Aortic Aneurysms and Dissection (TAAD1)	Dianna Milewicz	<a href="#">607087</a>	<a href="#">NISC</a>	Assigned
Mendelian	Paroxysmal Kinesigenic Dyskinesia (PKD)	Louis Ptacek	<a href="#">118800</a>	<a href="#">WUGSC</a>	Assigned
Mendelian	Atrial Fibrillation, Dominant (ATFB3)	Calum MacRae	<a href="#">609908</a>	<a href="#">RI-MIT</a>	Assigned
Allelic Spectrum	Age-Related Macular Degeneration	Goncalo Abecasis			Not Assigned
Allelic Spectrum	Diabetes	Michael Boehnke		<a href="#">NISC</a>	Assigned
Allelic Spectrum	Cardiovascular Disease/Diabetes	Eric Boerwinkle			Not Assigned
Allelic Spectrum	Metabolic Syndrome	Nelson Freimer		<a href="#">WUGSC</a>	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		<a href="#">RI-MIT</a>	Assigned
Allelic Spectrum	Tetralogy of Fallot	Christine Seidman			Not Assigned
Allelic Spectrum	Schizophrenia	Patrick Sullivan		<a href="#">BCM-HGSC</a>	Assigned

<http://www.genome.gov/20019648>

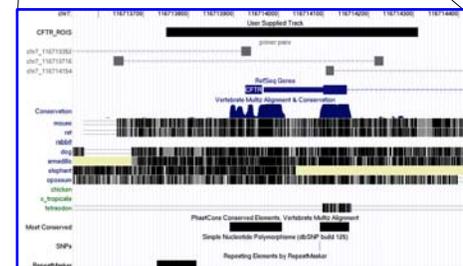
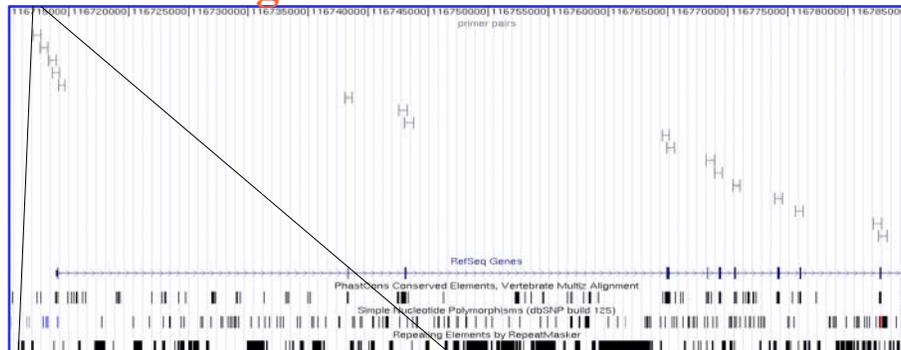
48

# 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

chr10:87,797,313-93,807,182

UCSC Gene Transcripts Based on RefSeq, UCSC Genes, and Comparative Genomics

chr10:87,797,313-93,807,182

Deletions from Microarray Analysis (CGP030)

Deletions from Microarray Analysis (Lafuze)

Deletions from Microarray Analysis (Ludman)

Deletions from Microarray Analysis (McCarton11)

Copy Number Polymorphisms from BWH (Sebat)

Copy Number Polymorphisms from BWH (Chen)

Structural Variation Identified by Fosmid (Tuzun)

Human Chromosome Self A1 (pansat)

51

## Primer Ordering and Tracking

3-D Barcode Order Form

Date: Thu Jun 22 17:07:38 2006  
 Customer: Keith Wetherby  
 Organization: NSC-NHGR/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# 420095240  
 Shipping Address: 3625 Fishers Lane, Room 5D-168  
 Bethesda, MD 20892  
 Billing Address: 3625 Fishers Lane, Room 5D-205  
 NIH 9439 Bethesda, MD 20892

Order Processing Details

Synthesis Scale: 8.01 used for all oligos in this order  
 Purity: HPLC (used for all oligos)  
 Method of Shipping: Lyophilized  
 Please Enter Additional Comments for Order Here: Samples should be in 1.5 ml tubes only

Number Oligo Names/Max of 11 characters

Seq#	1001740FOR.1	TGTAAAACGACGGCCAGT
1	1001741FOR.1	TGTAAAACGACGGCCAGT
3	1001742FOR.1	TGTAAAACGACGGCCAGT
4	1001743FOR.1	TGTAAAACGACGGCCAGT
5	1001744FOR.1	TGTAAAACGACGGCCAGT
6	1001745FOR.1	TGTAAAACGACGGCCAGT
7	1001746FOR.1	TGTAAAACGACGGCCAGT

found 49 entries

DBID	Name	Old Name	UCSC
1710	1001710	1003182	UCSC
1896	1001696	1003154	UCSC
1702	1001702	1003166	UCSC
1738	1001738	chr10_42882543	UCSC
1732	1001732	chr10_42883507	UCSC
1738	1001738	chr10_42920248	UCSC
1703	1001703	1003168	UCSC
1895	1001695	1003152	UCSC
1892	1001692	1003146	UCSC
1701	1001701	1003164	UCSC
1718	1001718	1003192	UCSC

found 41 entries

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

ROI ID	Location	Comment	Length	Amplifiers	Amplifier Design Coverage
2521	chr10:42786079-42786268	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	chr10:42803632-42803887	chr10_RET	256	2	100.0%
2526	chr10:42804019-42804428	chr10_RET	410	3	100.0%
2527	chr10:42805042-42805161	chr10_RET	120	1	100.0%

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

# Exploring the data

Projects

Amplifiers

ROIs

Primer Ordering

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

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

found 141 entries

Individual ID	Individual Count	Total Traces	Processed Traces	Number Amplifiers
41	CFTR_1	48	48	3
42	CFTR_16	48	48	3
43	CFTR_160	50	50	3
44	CFTR_161	22	22	3
45	CFTR_162	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 projects and progress overview

Individual 47

Note: CFTR\_104

Project: CFTR Resequencing

Attempted Amplifiers: 13

Successful Amplifiers: 12

Attempted Traces: 28

Successful Traces: 24

Median 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	Coverage
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	Amplifier 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

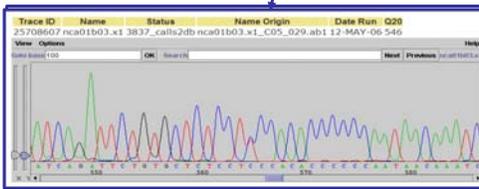
Individual	Alleles	Score	Trace	Trace Info	Strand
Hap_05	C/C	99	25822169	53129	-1
Hap_05	C/C	99	25821785	53137	1
HAPMAP_03	C/C	99	26204656	53153	-1
HAPMAP_03	C/C	99	25938327	53169	-1
HAPMAP_03	C/C	99	25936695	53127	1
HAPMAP_03	C/C	99	26202832	53134	1
AARS_8	C/C	99	25938363	53163	-1
AARS_8	C/C	99	25936731	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:318

ROI Location: chr1:210544826-216541138

Task: 8: mchick: wgs ( 238 chr + 431 xpa + 5.462 CPU)

Individual	Allele Name	Percent Coverage	Expected Count	Actual Count	Callotype Count	95% & 99% Coverage	Depth & 95% Reads Covered
105	HAP_05	100.0%	3	3	1	100.0%	210
201	HAPMAP_03	100.0%	4	4	1	100.0%	210
194	AARS_8	100.0%	4	4	1	100.0%	210
193	AARS_7	100.0%	4	4	1	100.0%	210
192	AARS_6	100.0%	4	4	1	100.0%	210
191	AARS_4	100.0%	4	4	1	100.0%	210
190	AARS_3	0%	4	0	0	0.0%	0
327	AARS_24	100.0%	4	4	1	100.0%	210
328	AARS_23	100.0%	4	4	1	100.0%	210



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.

55

MAJOR	C	C	T	T	G	G	C	C	G	G	C	C
MINOR	T	T	G	C	T	T	T	T	T	T	A	T
POLYID	2717	2716	2714	2721	2724	2719	2718	2720	2725	2720	2723	2727
CHR	chr16											
POSITION	16076029	16076109	16076284	16077633	16077978	16077978	16079069	16079069	16078116	16078116	16078272	16078348
AMPLIMER_ID	295	295	295	297	298	297	297	298	298	297	298	299
DBSNP	rs39821			rs148390	rs148390	rs148390	rs148390	rs39826	rs39826	rs39827		
CROSS QUERIES	INTRONIC			INTRONIC								
Hap_01	CC	CC	TC									
ABCO01_1	CC	CC	TC									
ABCO01_2	CC	CC	TC									
ABCO01_3	CC	CC	TC									
ABCO01_4	CC	CC	TC									
ABCO01_5	CC	CC	TC									
ABCO01_6	TC											
ABCO01_7	CC	CC	TC									
HAPMAP_04	CC	CC	TC									
Hap_06	CC	CC	TC									
ABCO01_19	CC	CC	TC									
ABCO01_21	CC	CC	TC									
ABCO01_22	CC	CC	TC									
ABCO01_24	CC	CC	TC									
ABCO01_25	CC	CC	TC									
ABCO01_26	CC	CC	TC									
ABCO01_28	CC	CC	TC									
ABCO01_29	CC	CC	TC									
ABCO01_30	CC	CC	TC									
ABCO01_32	TC											
ABCO01_33	CC	CC	TC									
ABCO01_34	TC											
ABCO01_35	CC	CC	TC									

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.

56

## Some of the challenges of variation detection



Heterozygous DIPs

“Dye blob”

Detection saturation

57

## Future of Medical Sequencing

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

58

## 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

59

## 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

60

  
**454 FLX**  
**Pyrosequencing**

  
**Genome Analyzer  
(Solexa)**  
**Sequencing by synthesis**

  
**SOLiD**  
**Ligation-based extension**

61

## Platform Comparisons

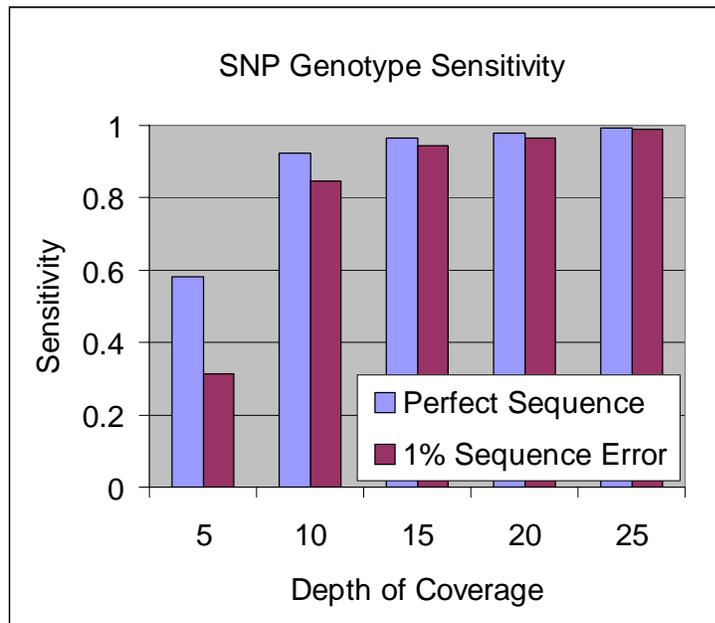
Criterion	ABI 3730	Roche 454	Illumina	AB Solid
Sequencing chemistry	Big dye ddNTPs	Pyrosequencing	Sequencing by synthesis	Ligation-based sequencing
Amplification approach	Linear PCR	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired ends/separation	Yes/variable	Yes/3kb	Yes/200bp	Yes/3kb
Time/run (bases/run)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

62

## 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)

63



64

## Example of short read sequence alignment

```

Reference:  GAATAGCCTAAATATGGTAAAATATTTTITTCATTATATCTTTAGATTCATCAATTTTATATAAATGAAAATGGAAAGCAGCTTAAGTCTCAC
Called:      * ..... | 91410
F>SILVA-EASL_04 3 145 239 379 -2 AAATAGCCTAAATATGGTAAAAT
F>SILVA-EASL_04 3 169 165 775 -1 AAATAGCCTAAATATGGTAAAAT
F>SILVA-EASL_04 3 175 993 578 -2 AATAGCCTAAATATGGTAAAATCTT
R>SILVA-EASL_04 3 179 524 423 -2 AAAGCCTAAATATGGTAAAATATT
R>SILVA-EASL_04 3 70 304 976 -1 ATAGCCTAAATATGGTAAAATATT
R>SILVA-EASL_04 3 70 338 440 -2 ATAGCCTAAATATGGTAAAATATT
P>SILVA-EASL_04 3 44 927 371 -1 AGCCTAAATATGGTAAAATATTTT
F>SILVA-EASL_04 3 189 48 723 -2 GCTTAAATATGGTAAAATATTTT
F>SILVA-EASL_04 3 51 253 942 -1 GCTTAAATATGGTAAAATATTTT
R>SILVA-EASL_04 3 193 175 960 -1 TAAATATGGTAAAATATTTTTCCL
F>SILVA-EASL_04 3 163 723 431 -2 AAATATGGTAAAATATTTTTCAT
F>SILVA-EASL_04 3 120 813 254 -1 AATATGGTAAAATATTTTTCATTA
R>SILVA-EASL_04 3 9 981 462 -1 ATATGGTAAAATATTTTTCATTA
P>SILVA-EASL_04 3 144 742 304 -1 ATGGTAAAATATTTTTCATTATA
F>SILVA-EASL_04 3 170 984 902 -1 ATGGTAAAATATTTTTCATTATA
R>SILVA-EASL_04 3 36 40 188 -1 GCTTAAATATTTTTCATTATATC
F>SILVA-EASL_04 3 52 196 922 -2 TAAAATATTTTTCATTATATCTT
F>SILVA-EASL_04 3 97 106 724 -2 AAATATTTTTCATTATATCTT
R>SILVA-EASL_04 3 57 309 602 -2 AAATATTTTTCATTATATCTT
F>SILVA-EASL_04 3 26 502 310 -2 ATTTTTCATTATATCTTACATT
R>SILVA-EASL_04 3 45 133 458 -1 ATTTTTCATTATATCTTACATT
R>SILVA-EASL_04 3 68 901 127 -2 ATTTTTCATTATATCTTACATT
R>SILVA-EASL_04 3 118 824 15 -1 TTTTTCATTATATCTTACATCG
R>SILVA-EASL_04 3 36 656 711 -2 attatctcttaccctctgcaatttt
F>SILVA-EASL_04 3 8 208 936 -2 atctctcttaccctctgcaattttt
R>SILVA-EASL_04 3 135 793 41 -2 CTTTACATTCATCAATTTTATTAT
F>SILVA-EASL_04 3 11 224 677 -1 TTACATTCATCAATTTTATTATTA
R>SILVA-EASL_04 3 143 553 701 -2 TTACATTCATCAATTTTATTATTA
R>SILVA-EASL_04 3 123 521 120 -2 CATTACATCAATTTTATTATTA
F>SILVA-EASL_04 3 81 763 828 -1 TCTCAATTTTATTATTAATATG
F>SILVA-EASL_04 3 60 115 315 -1 TTTTATTATAAATGAAAATGG
F>SILVA-EASL_04 3 512 601 -1 TTTTATTATAAATGAAAATGG
F>SILVA-EASL_04 3 2 17 229 -2 TTATTATAAATGAAAATGG
R>SILVA-EASL_04 3 186 362 486 -2 ATATAATGAAAATGCAAGCCAG
R>SILVA-EASL_04 3 122 742 949 -2 taataatgaaaatggaagacagc
R>SILVA-EASL_04 3 84 695 956 -2 aataatgaaaatggaagacagc
F>SILVA-EASL_04 3 85 745 -2 AATAAATAAATGAAAATGCAAGCC
F>SILVA-EASL_04 3 106 574 300 -2 AATGAAAATGAAAATGCAAGCC
F>SILVA-EASL_04 3 79 874 873 -2 ATGAAAATGAAAATGCAAGCC
F>SILVA-EASL_04 3 168 384 873 -2 ATGAAAATGAAAATGCAAGCC

```

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



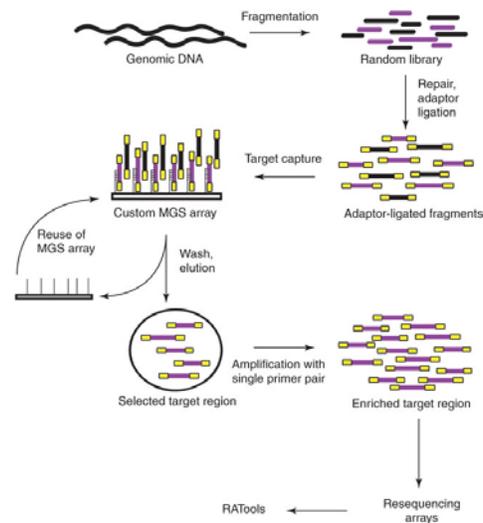
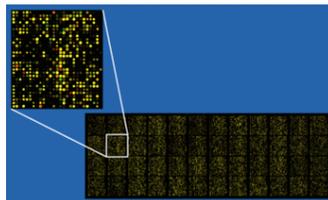
# Targeted Genomic Selection

- 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

69

## Microarray Direct Capture

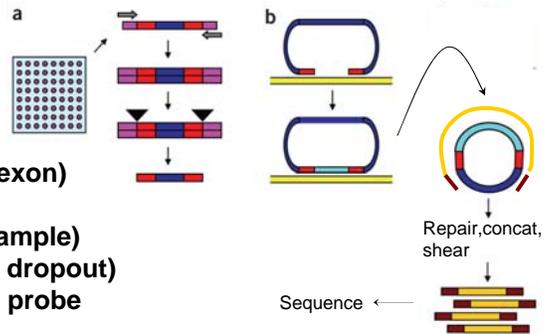
- 385k features / chip (6MB seq.)
- 24-30k exons / chip
- 55-85% specificity (seq in ROI)
- 12-50% total ROI seq coverage\*
- Exon coverage 40-78% (22-60% dropout)
- Non-uniform seq. depth
- 20 ug DNA input



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

70

## Molecular Inversion Probe Capture

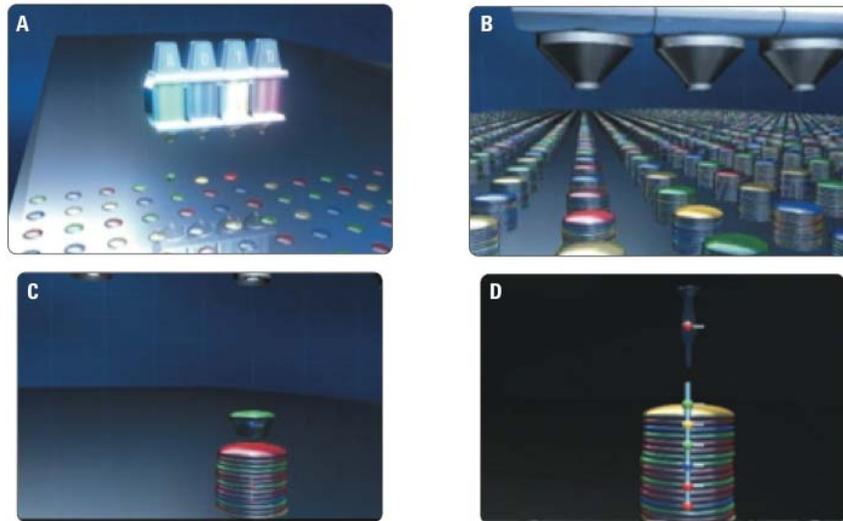


- 55,000 capture oligos (1 / exon)
- 6.7 Mb total seq.
- Specificity = 98.6% (small sample)
- Exon coverage = 91% (9% dropout)
- Each exon targeted with 1 probe
- 750 ng - 1.5 ug DNA input
- Highly non-uniform seq. coverage (several logs) - but consistent
- Het calls - 96% sens.

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

71

## High Throughput Synthesis Of Long Oligo Libraries



 Agilent Technologies

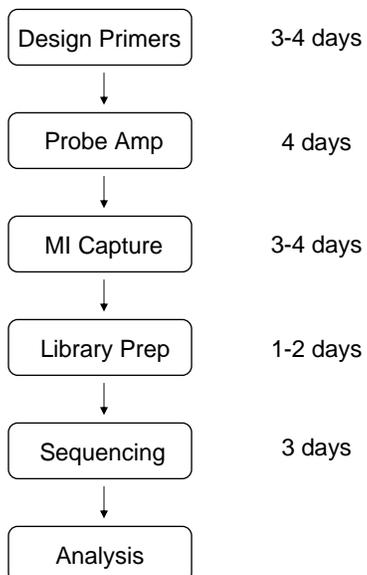
72

## Molecular Inversion Probe Capture



73

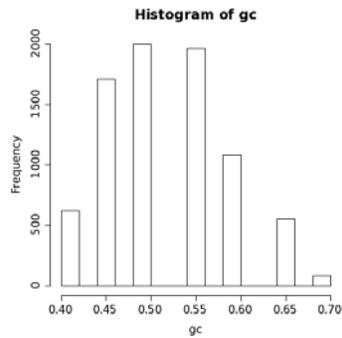
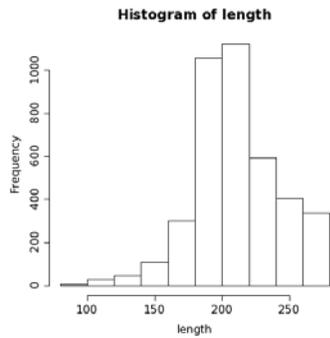
## Experimental Design - Workflow



74

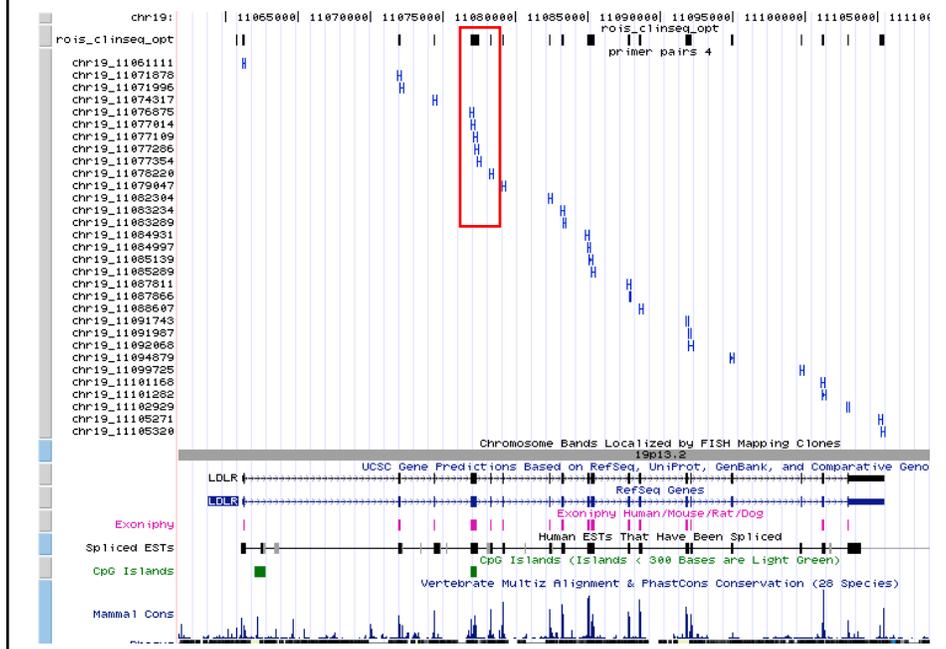
## Experimental Design - Primers

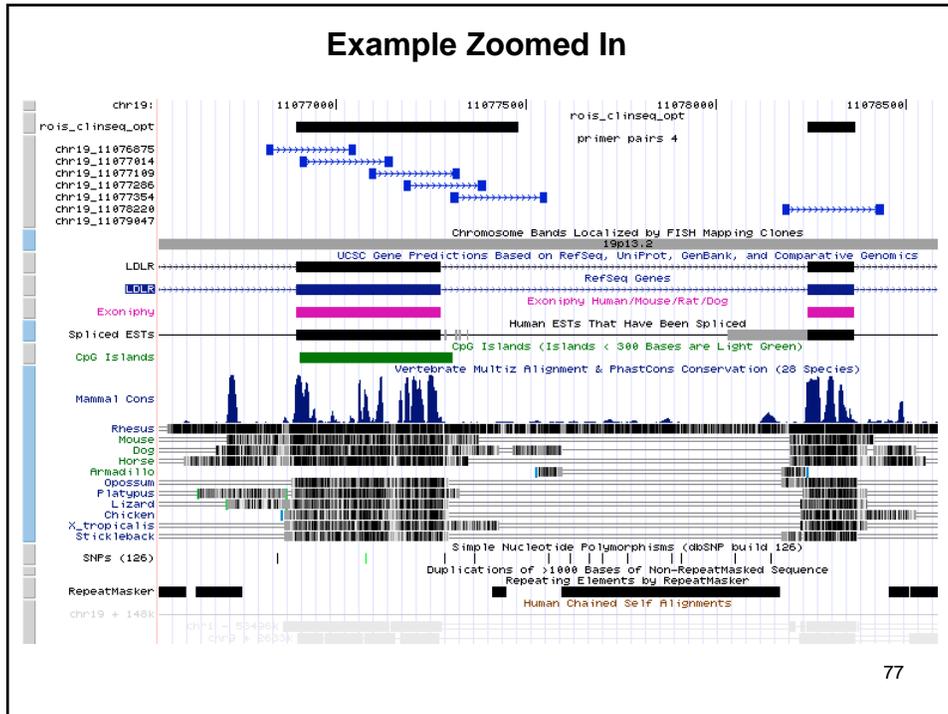
- 4000 probes - (492kbp)
- Use primer\_tile to design - similar to PCR primers
- Capture Region Criteria
  - length of region: 90-280bp, 200 optimal
  - GC% in targeting pairs: 45-65%, then 40-70%
  - minimize non-specific targeting pairs using ePCR
  - no Nt.AIwI or Nb.BsrDI restriction sites in targeting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



75

## Example Region





- ### 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
- 78

# 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/>

79

# 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.

80

# 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.

81

# 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.

82

## 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](http://genome.perlegen.com/browser/index_v2.html): Perlegen's HapMap

<http://www.genome.gov/25521748> : HGSV