Human Genetic Variation and Association Studies

Jacques Fellay, MD
SNF Professor

Global Health Institute
EPFL School of Life Sciences

Institute of Microbiology
University of Lausanne / CHUV

We Are All Different

We Are All Different

About 2% of people have two copies of APOE4 and are very likely to succumb to Alzheimer’s disease

We Are All Different

About 2% of people have two copies of APOE4 and are very likely to succumb to Alzheimer’s disease

About 1% of us have two copies of a small deletion in CCR5 and are largely immune to infection by the HIV virus

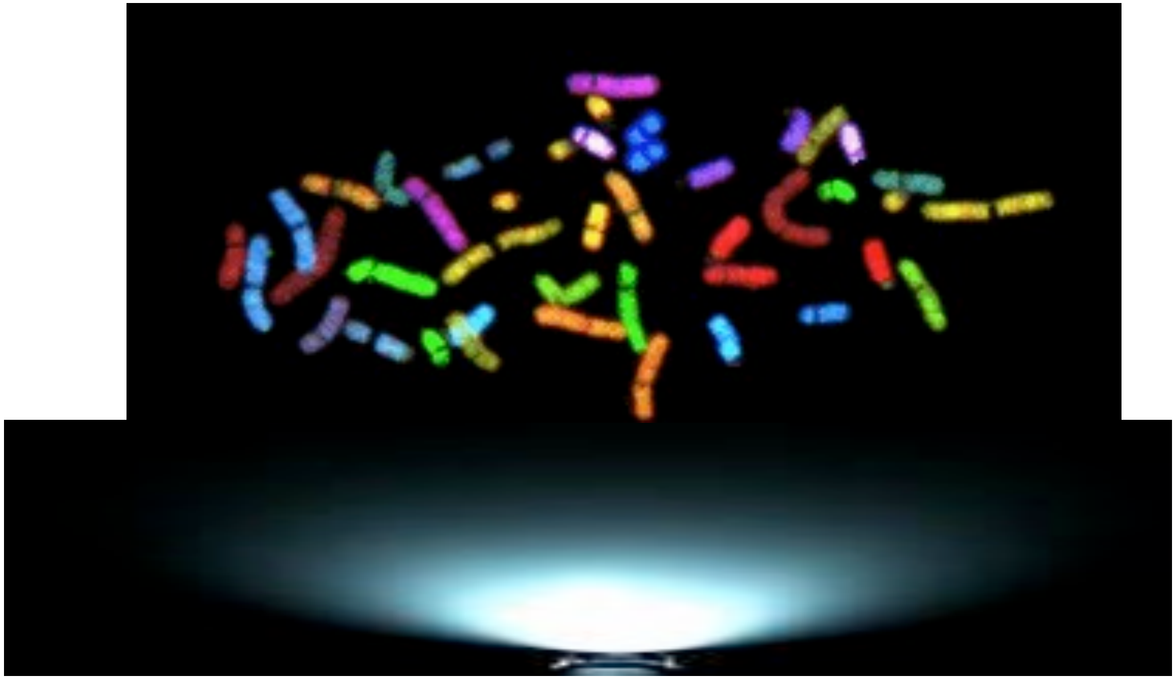
We Are All Different

About 2% of people have two copies of APOE4 and are very likely to succumb to Alzheimer’s disease

About 1% of us have two copies of a small deletion in CCR5 and are largely immune to infection by the HIV virus

And about 7% do not make any functional CYP2D6 enzyme and therefore codeine provides no pain relief

These examples come from looking at only the tiniest fraction of our genome



It is now possible to scan the *whole genome* to find the genetic determinants of key differences amongst people

Overview

- Types of human genetic variation

- Mapping approaches
  - GWAS
  - Sequencing

- Real life examples

Different forms of genetic variation

Different forms of genetic variation

- -single nucleotide variants
  - -3-4 million per individual

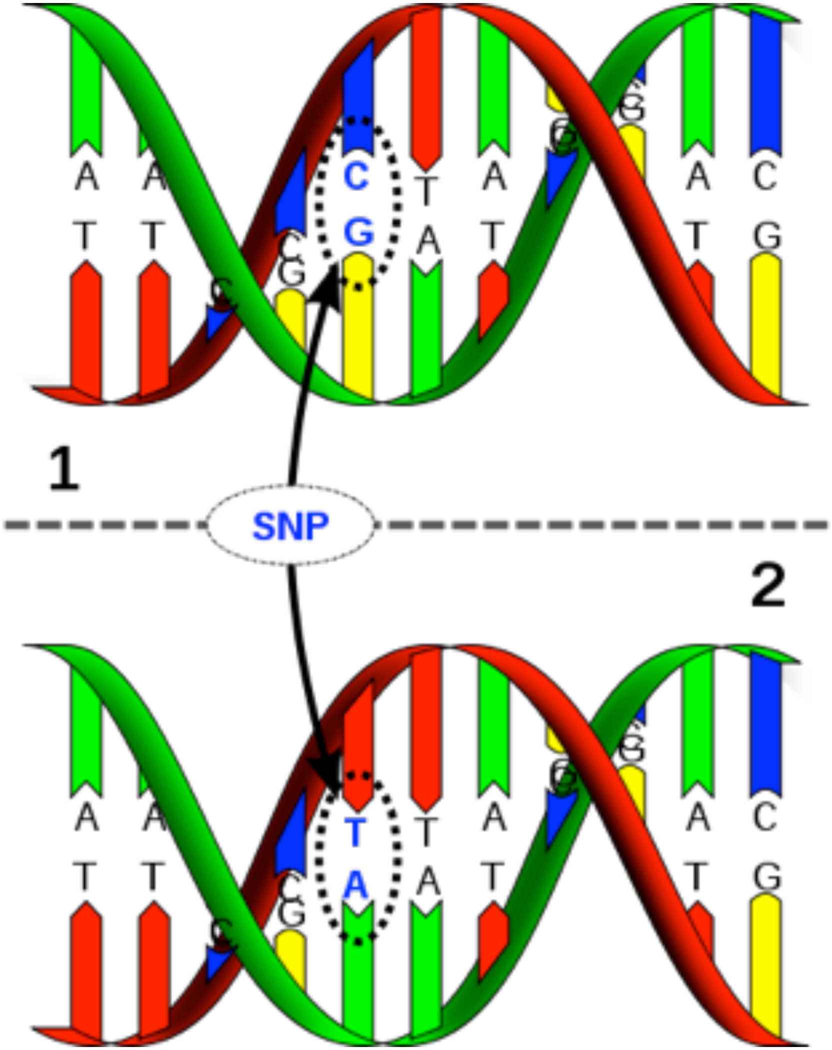
- -multiple nucleotides variants
  - -greater content than single site changes

Different forms of genetic variation

-single nucleotide polymorphisms (or SNPs)

functional?
- -missense
- -non-sense
- -splice site

non-functional?
- -silent
- -intronic
- -intergenic

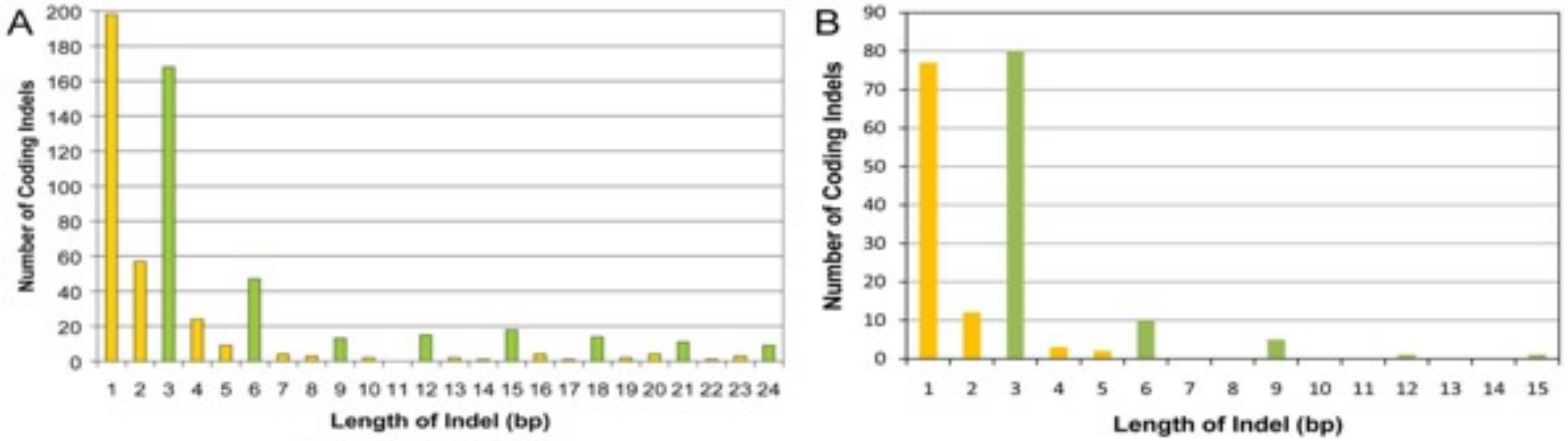


Different forms of genetic variation

-single nucleotide polymorphisms (or SNPs)

-small insertions or deletions (indels)

-coding/non-coding



J. C. Venter's Genome
(Figure from Ng et al. PLoS Genetics 4(8): e1000160)

This study
(individual genome average)

Different forms of genetic variation

-single nucleotide polymorphisms (or SNPs)

-small insertions or deletions (indels)

-short tandem repeats/microsatellites

- -repeat of 2, 3, 4 or more nucleotides
- -10-100x
- -highly polymorphic
- -error during replication (slippage)

Trinucleotide repeat diseases

14 known diseases

-9 due to glutamine repeats (CAG trinucleotide)
-neurodegenerative disease (polyglutamine disease)
-neuronal decay
-spinocerebellar ataxias and Huntington’s disease

Trinucleotide repeat diseases

14 known diseases

-9 due to CAG trinucleotide=Glutamine
-neurodegenerative disease (polyglutamine disease)
-neuronal decay
-Spinocerebellar ataxias and Huntington's disease

Huntington's disease trinucleotide repeats

-tract of <28	=normal
-tract of 28 to 35	=intermediate
-tract of 36-40	=reduced penetrance/affected
-tract of >40	=full penetrance/affected

anticipation – tract expands with successive generations leading to earlier age of onset and more severe disease.

Different forms of genetic variation

-single nucleotide polymorphisms (or SNPs)
-small insertions or deletions (indels)
-short tandem repeats/microsatellites
-retrotransposons (RNA intermediate)

LINE -long interspersed repetitive elements
-makes copies
-17% of genome

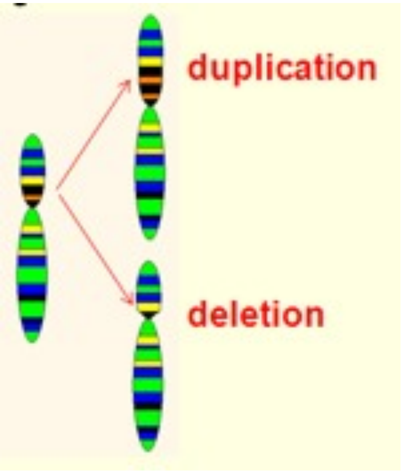
SINE -short interspersed repetitive elements
-alu sequence (around 300 bp)
-over 10% of genome

LTR -long terminal repeats
-8% of genome
-still active?

Different forms of genetic variation

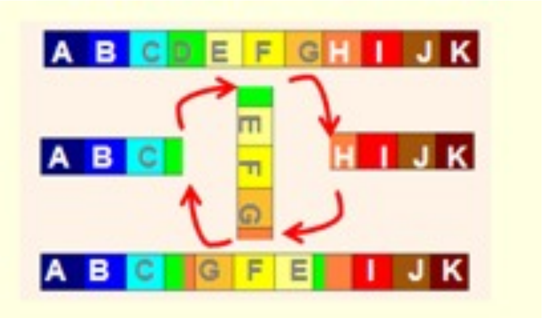
-single nucleotide polymorphisms (or SNPs)
-small insertions or deletions (indels)
-short tandem repeats/microsatellites
- retrotransposons (RNA intermediate)
-copy number variants (CNVs)
-deletions/duplications

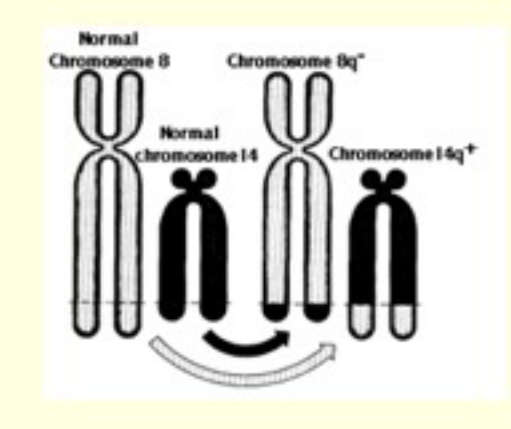
CNVs



Different forms of genetic variation

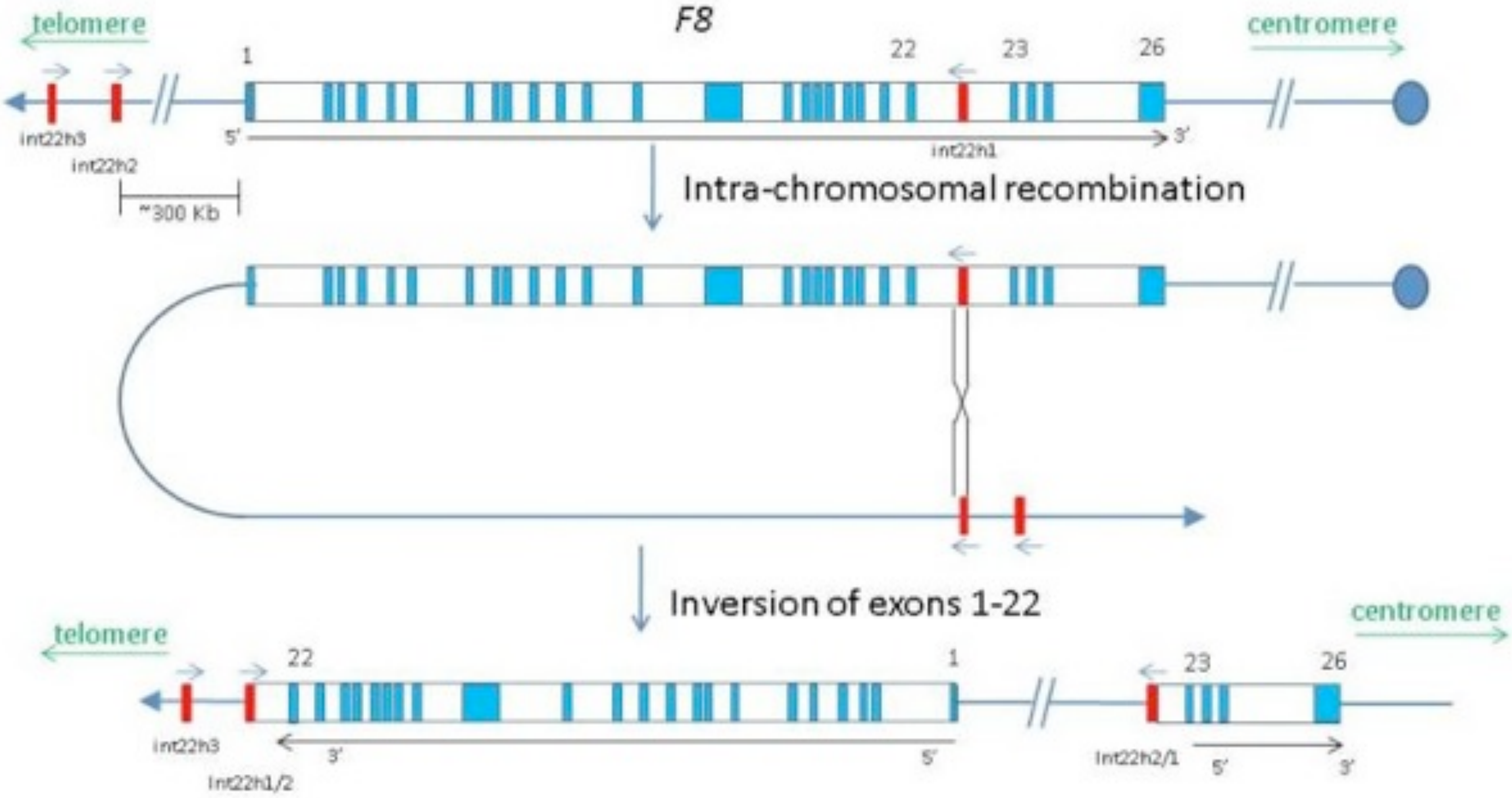
-single nucleotide polymorphisms (or SNPs)
-small insertions or deletions (indels)
-short tandem repeats/microsatellites
-transposable elements
-copy number variants (CNVs)
-large structural variation
    -inversions
    -translocations





Factor VIII gene inversions

-severe hemophilia A
-40% of individuals have a large 400kb inversion



Variants may be…

-common (>1%)
-rare (<1%)

-single family
-small region
-one population
-all populations

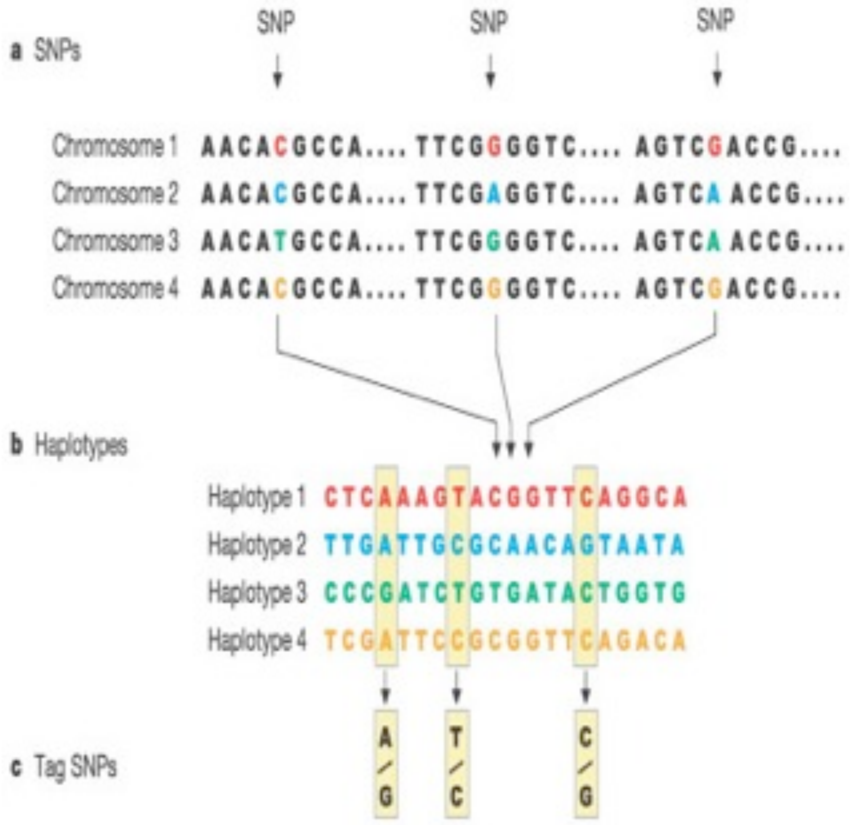
International HapMap Project

- Identify SNPs from 270 individuals
  - CEU: CEPH (Utah residents with ancestry from northern and western Europe) (30 trios)
  - CHB: Han Chinese in Beijing, China (45 individuals)
  - JPT: Japanese in Tokyo, Japan (45 individuals)
  - YRI: Yoruba in Ibadan, Nigeria (30 trios)

International HapMap Project

Identify SNPs from 270 individuals

- Establish haplotypes
- Identify Tagging SNPs


a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

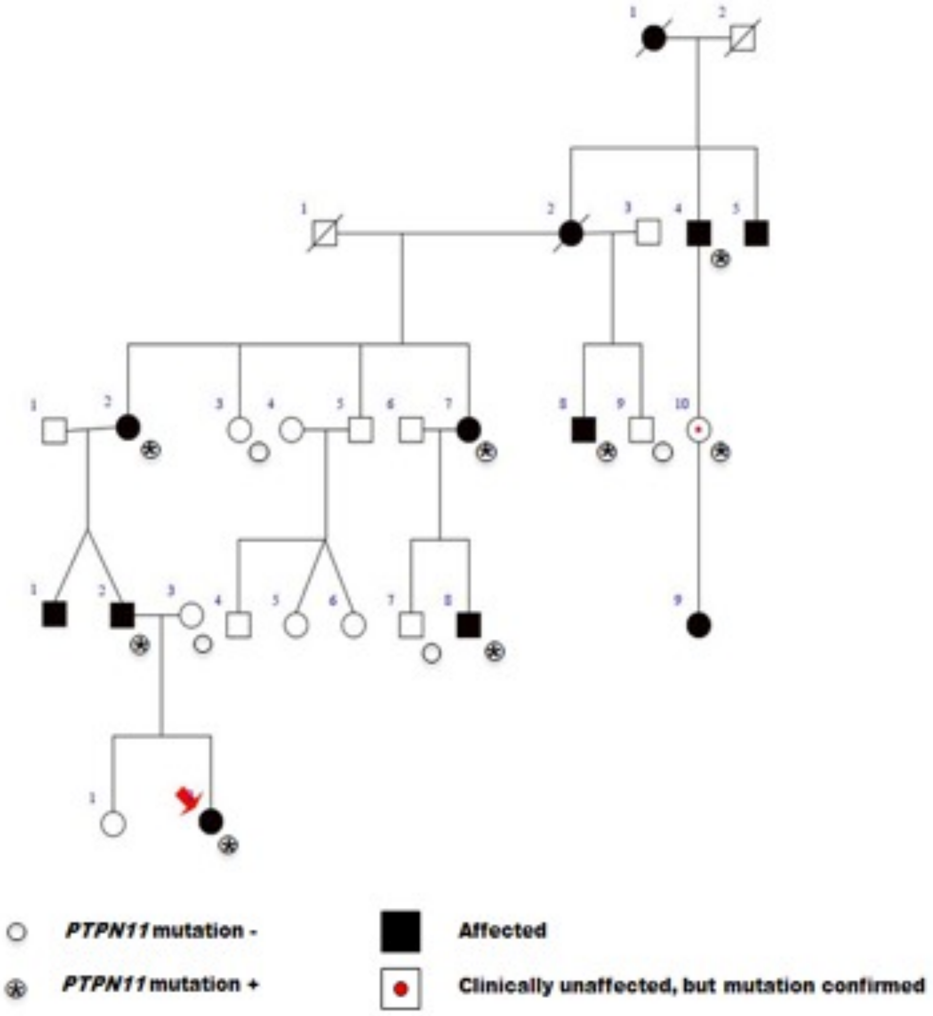
Finding gene variants

Mapping strategies

- Map based
  - Use a set of markers spread throughout the genome designed to capture most regions/ common variants
- Complete resequencing

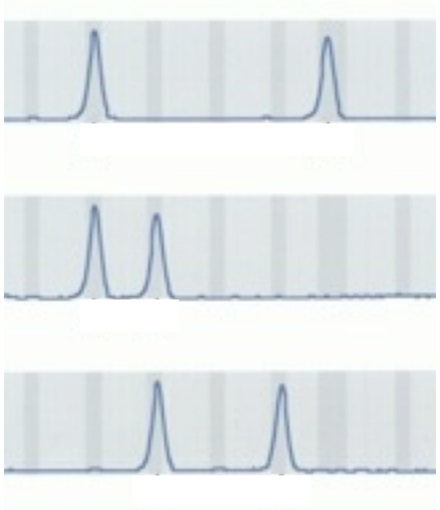
Mapping using genetic variation

linkage analysis using microsatellites



-genotype 300-400 markers for affected and unaffected individuals

-10 cM resolution = 10 Mb



Mapping using genetic variation

Genome-Wide Association Studies (GWAS), using SNPs

Goal: identify common variation associated with a specific phenotype/trait

GWAS Basic Strategy

- Large sample size (1000’s)
- Well defined phenotype
- Case/control or continuous phenotype
- Whole-genome genotyping
- Appropriate statistics

  -correction for multiple testing

  -population stratification

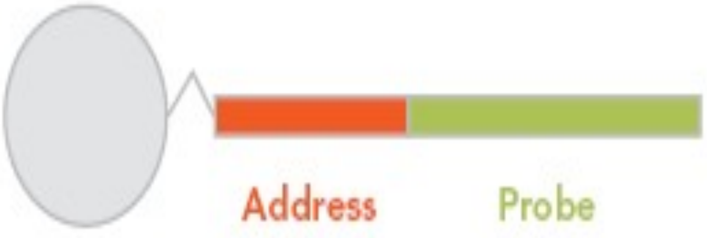
Whole Genome Genotyping

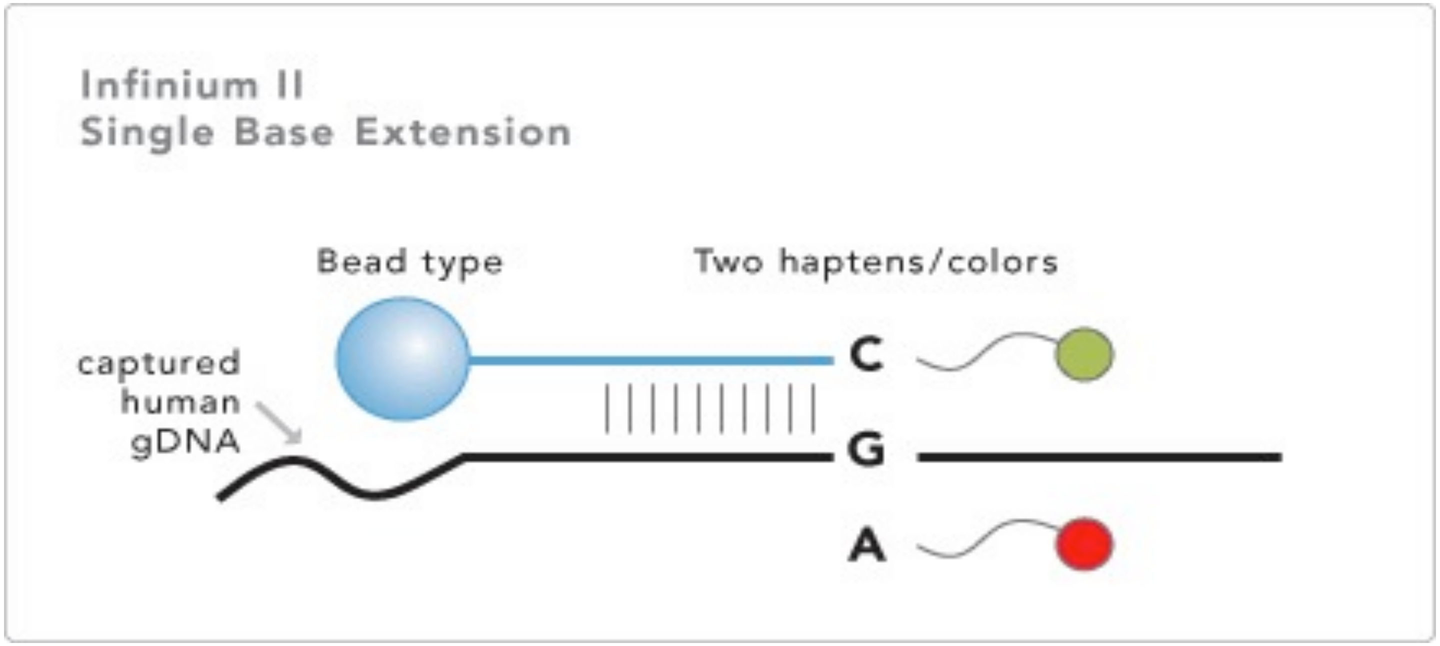
-10 million common variants in our genome

-Genotyping chips can assay over 2.5 million SNVs

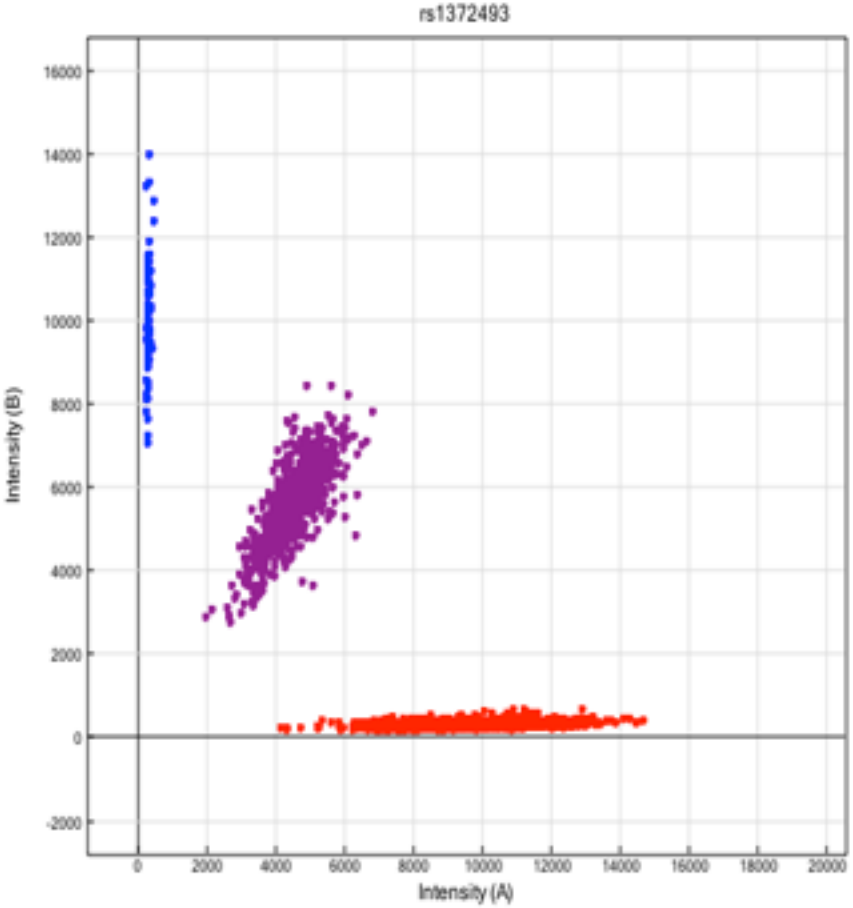
-Excellent coverage of common genetic variants for most populations

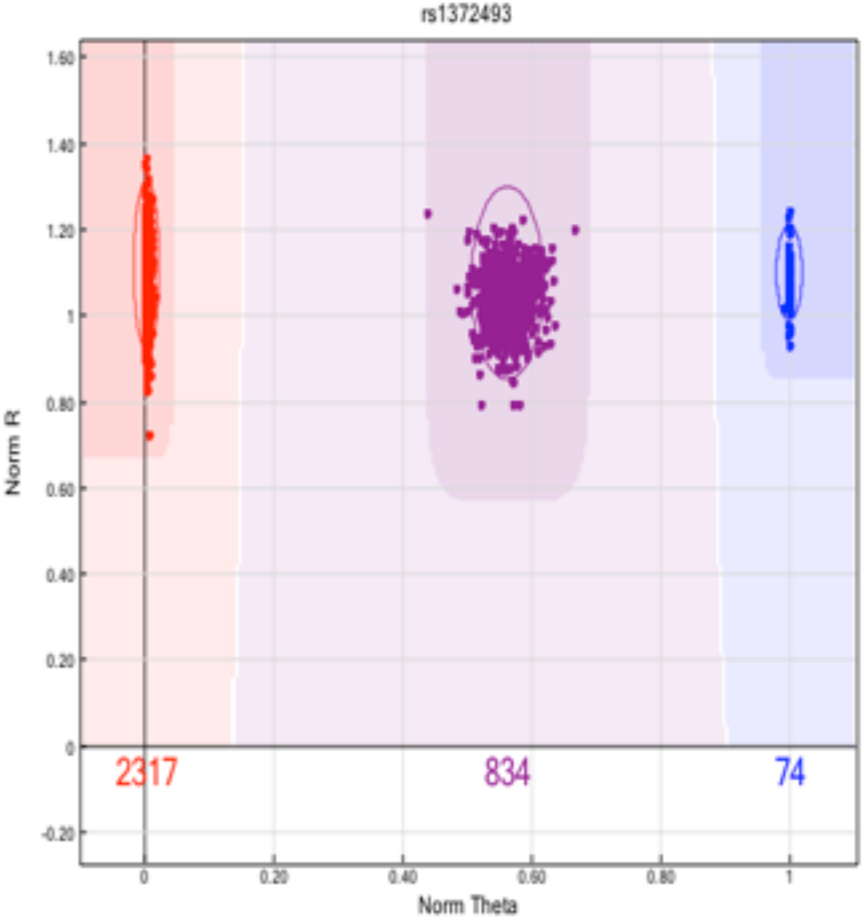
Illumina Infinium Assay

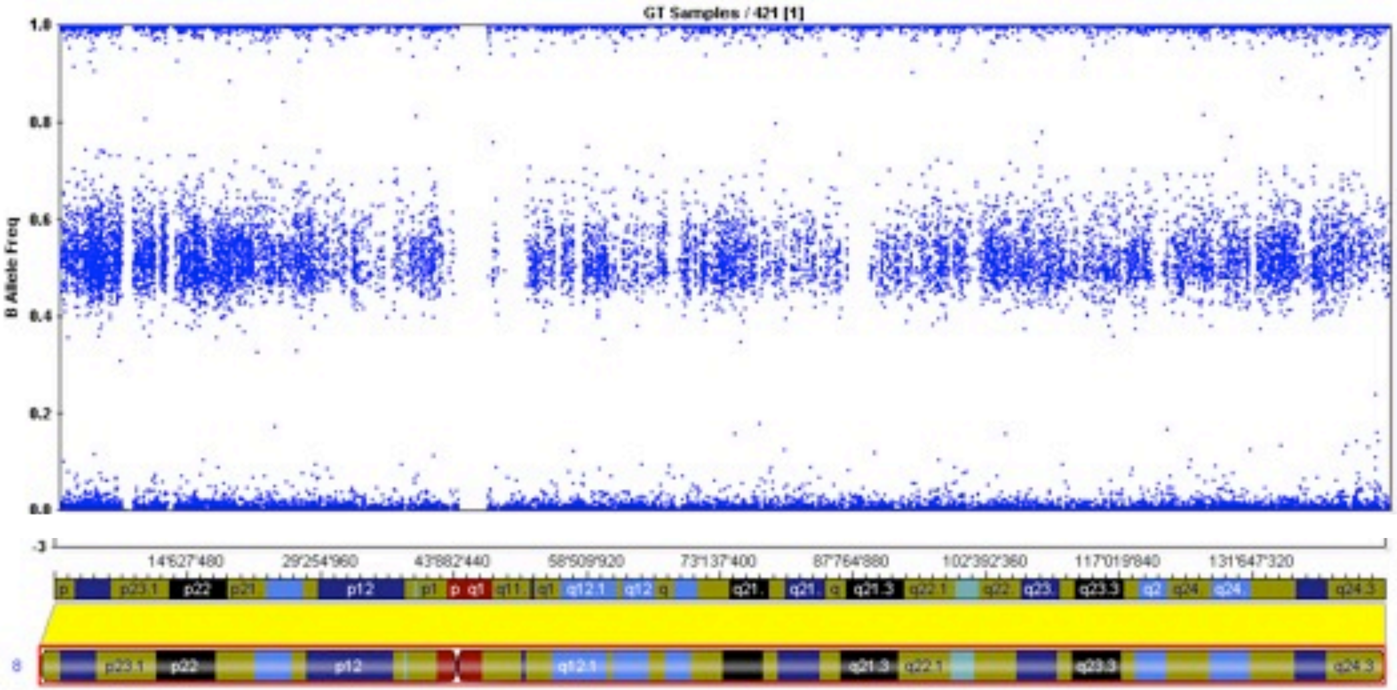




SNP output


rs1372493
Intensity (B)
16000
14000
12000
10000
8000
6000
4000
2000
0
-2000
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
Intensity (A)

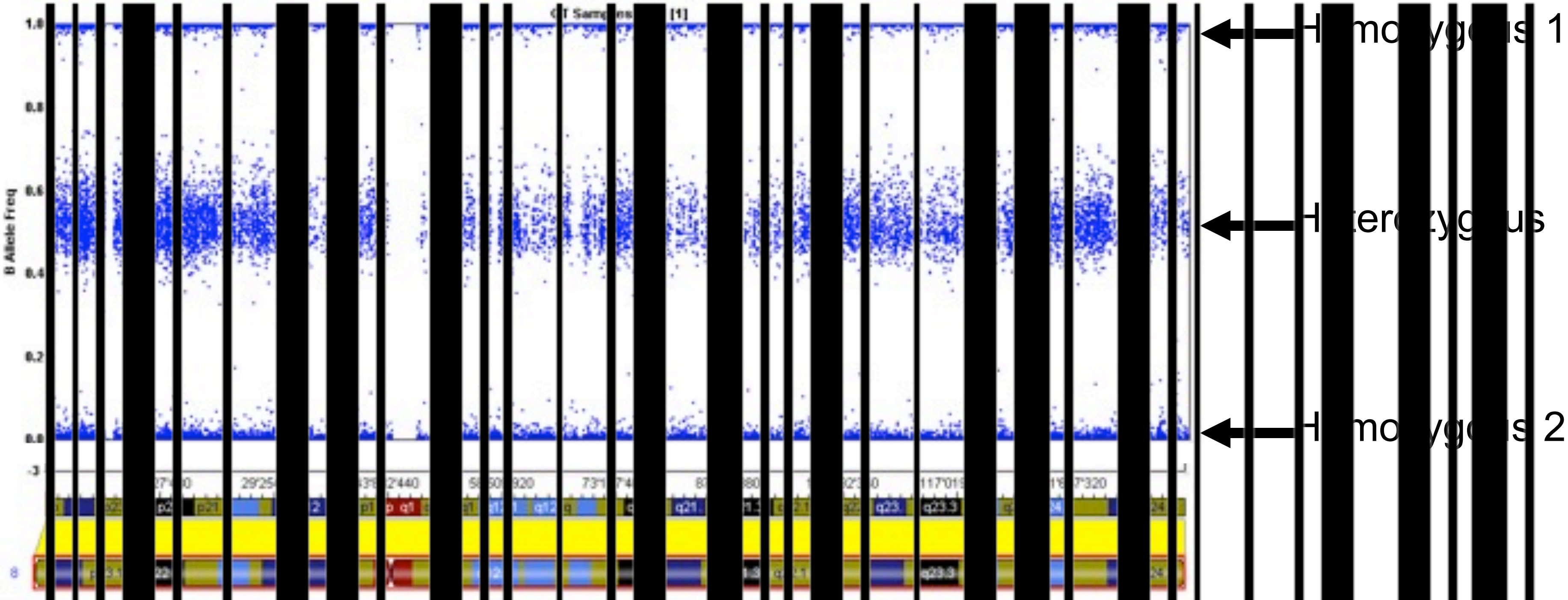

rs1372493
Norm R
1.60
1.40
1.20
1
0.80
0.60
0.40
0.20
0
-0.20
2317
834
74
0
0.20
0.40
0.60
0.80
1
Norm Theta



Homozygous 1

Heterozygous

Homozygous 2



H[redacted]mo[redacted]yg[redacted]us 1

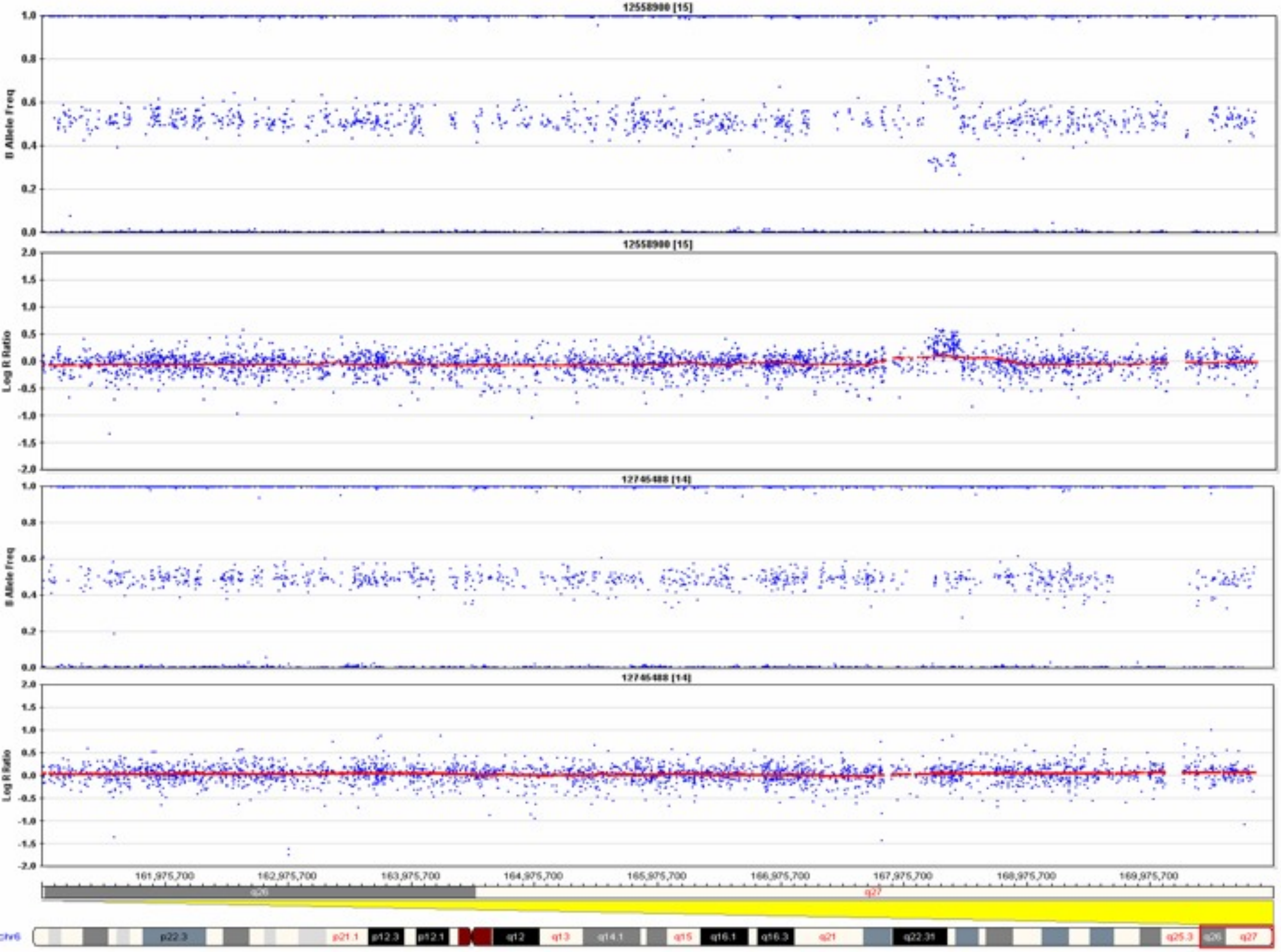
H[redacted]tero[redacted]zyg[redacted]us

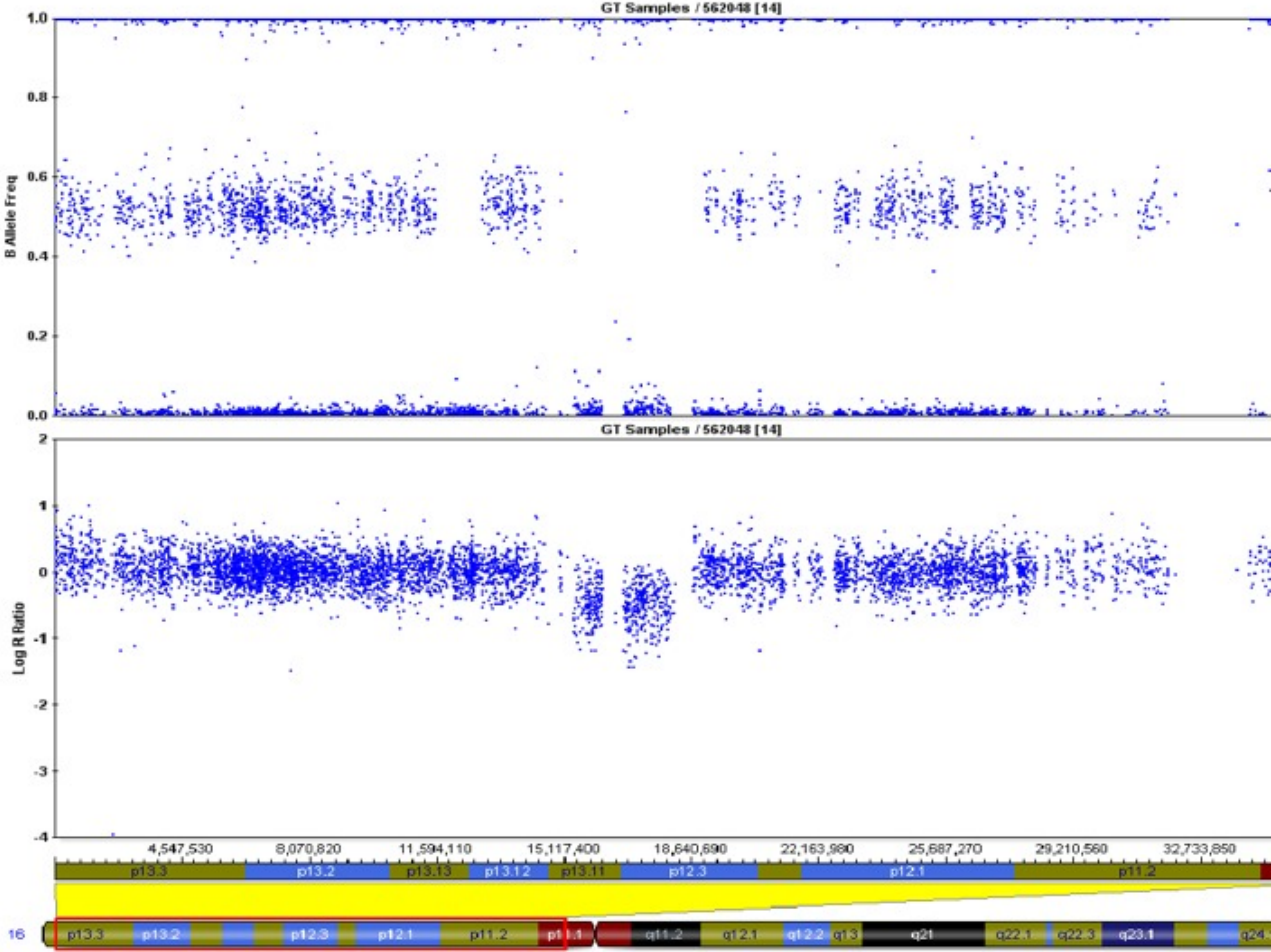
H[redacted]mo[redacted]yg[redacted]us 2

Copy Number Variation (CNVs)

-Many known common insertions/deletions identified throughout the genome

-Whole-genome genotyping arrays can detect larger CNV events (>50kb)


12558900 [15]
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
12558900 [15]
Log R Ratio
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
12745488 [14]
B Allele Freq
12745488 [14]
Log R Ratio
161,975,700
162,975,700
163,975,700
164,975,700
165,975,700
166,975,700
167,975,700
168,975,700
169,975,700
q26
q27
chr6
p22.3
p21.1
p12.3
p12.1
q12
q13
q14.1
q15
q16.1
q16.3
q21
q22.31
q25.3
q26
q27


GT Samples / 562048 [14]
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
GT Samples / 562048 [14]
Log R Ratio
2
1
0
-1
-2
-3
-4
4,547,530
8,070,820
11,594,110
15,117,400
18,640,690
22,163,980
25,687,270
29,210,560
32,733,850
p13.3
p13.2
p13.13
p13.12
p13.11
p12.3
p12.1
p11.2
16
p13.3
p13.2
p12.3
p12.1
p11.2
p11.1
q11.2
q12.1
q12.2
q13
q21
q22.1
q22.3
q23.1
q24

Common vs Rare Variation

- GWAS generally only have the power to detect common SNP variation

- Rare variation?
  - Genome Sequencing

Transition to Sequencing

Moore’s Law

sequencing throughput/cost

2007 2008 2009 2010 2011

single run 1Gb

single run 650 Gb

Next Generation Sequencing Setup

-Huge amount of data (terabytes!)

-Analysis computationally intensive

-Dedicated IT infrastructure

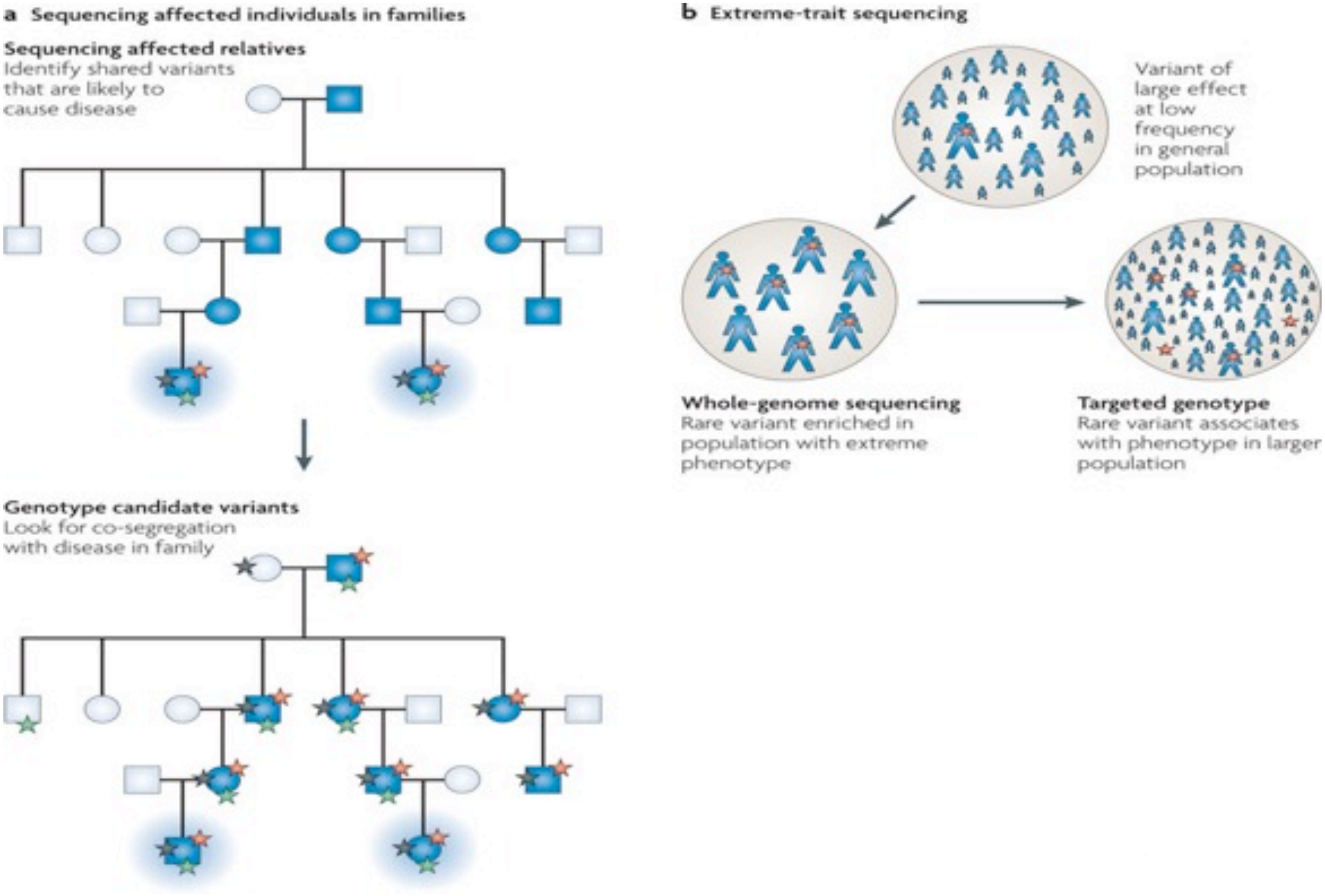
Sequencing Approach

1. Identify subjects

   -Extreme phenotype or family based
2. Sequence (50-100 individuals)
3. Align to reference and call variants
4. Compare to 100’s of sequenced controls

5. Follow-up genotyping in **larger** cohorts!

Sequencing Approach


a Sequencing affected individuals in families
Sequencing affected relatives
Identify shared variants that are likely to cause disease
Genotype candidate variants
Look for co-segregation with disease in family
b Extreme-trait sequencing
Variant of large effect at low frequency in general population
Whole-genome sequencing
Rare variant enriched in population with extreme phenotype
Targeted genotype
Rare variant associates with phenotype in larger population

Cirulli ET& Goldstein DB. *Nature Reviews Genetics*, 2010

Filtered FastQ sequence

Data from a single cluster/read (75bp)

```
@G:1:1:11:1079#0/1
TGATTGATTCCATTCCATTCCATTCCATTTCATTCCATTGCAATCCCTTCCAATCCATTCCATTCCATTCCATTC

+G:1:1:11:1079#0/1
`Xa^YO\_^a_`__`a__^a^a^_a``^_\`\\]``[XUGXXXXXWUTWWVWUSTXXPUWYYRVWYYYXZYXYWZ
```

*a finished genome will have over 1 billion reads

Analysis

- Alignment to reference genome
  - 3 billion bases
- Call variants
  - Single nucleotide variants (SNVs)
  - Small insertion/deletions (indels)
  - Structural Variants (SV/CNV)

Summary of a single human genome

SNVs	3.5 million
Premature stop	120
Stop loss	25
Non-synonymous	11,000
Synonymous	11,000
Essential splice site	100

indels	610,000
Frameshift	500
In-frame	900

Whole genome vs. exome sequencing

Exome

- -Coding regions
- -Cheaper/Faster
- -Uneven capture of both alleles
- -Incomplete capture of target region
- -Bias towards known biology

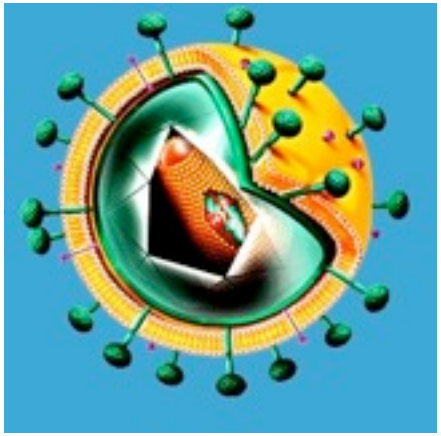
Genome

- -Complete sequence
- -Expensive/Throughput
- -IT issues (10 fold more data)

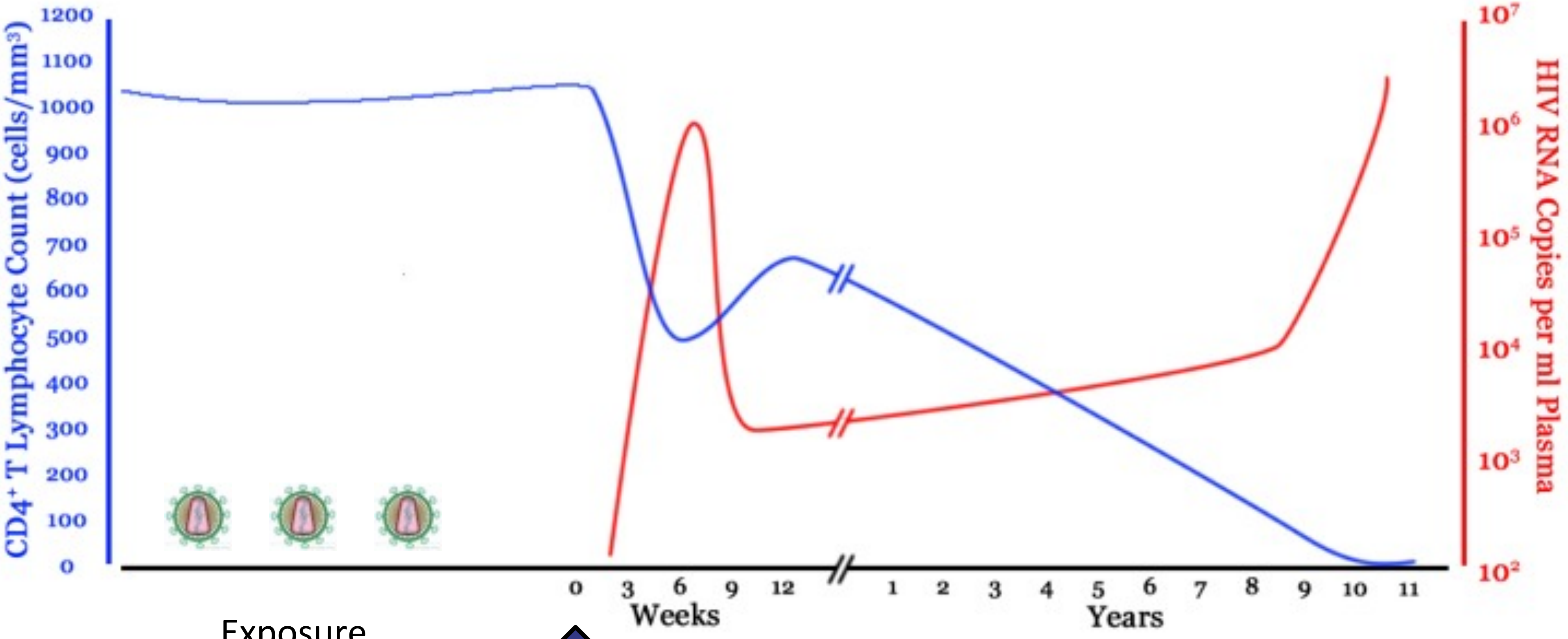
GWAS Examples

Host genetic differences contribute to variation in response to HIV

- ➢ Susceptibility to infection
- ➢ Natural history of disease
  - ○ Viral load
  - ○ Immunological progression
  - ○ AIDS events / death

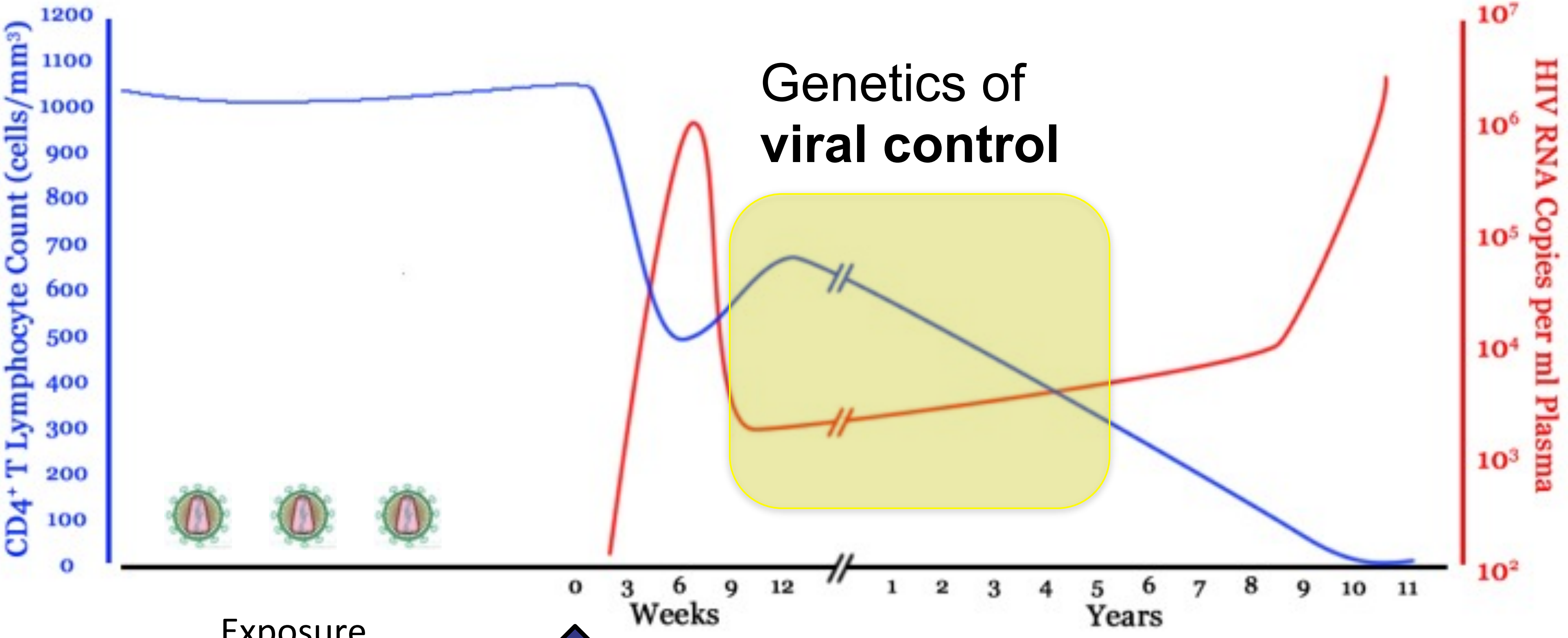


Host genetics of HIV disease



Exposure

Infection



Genetics of
viral control

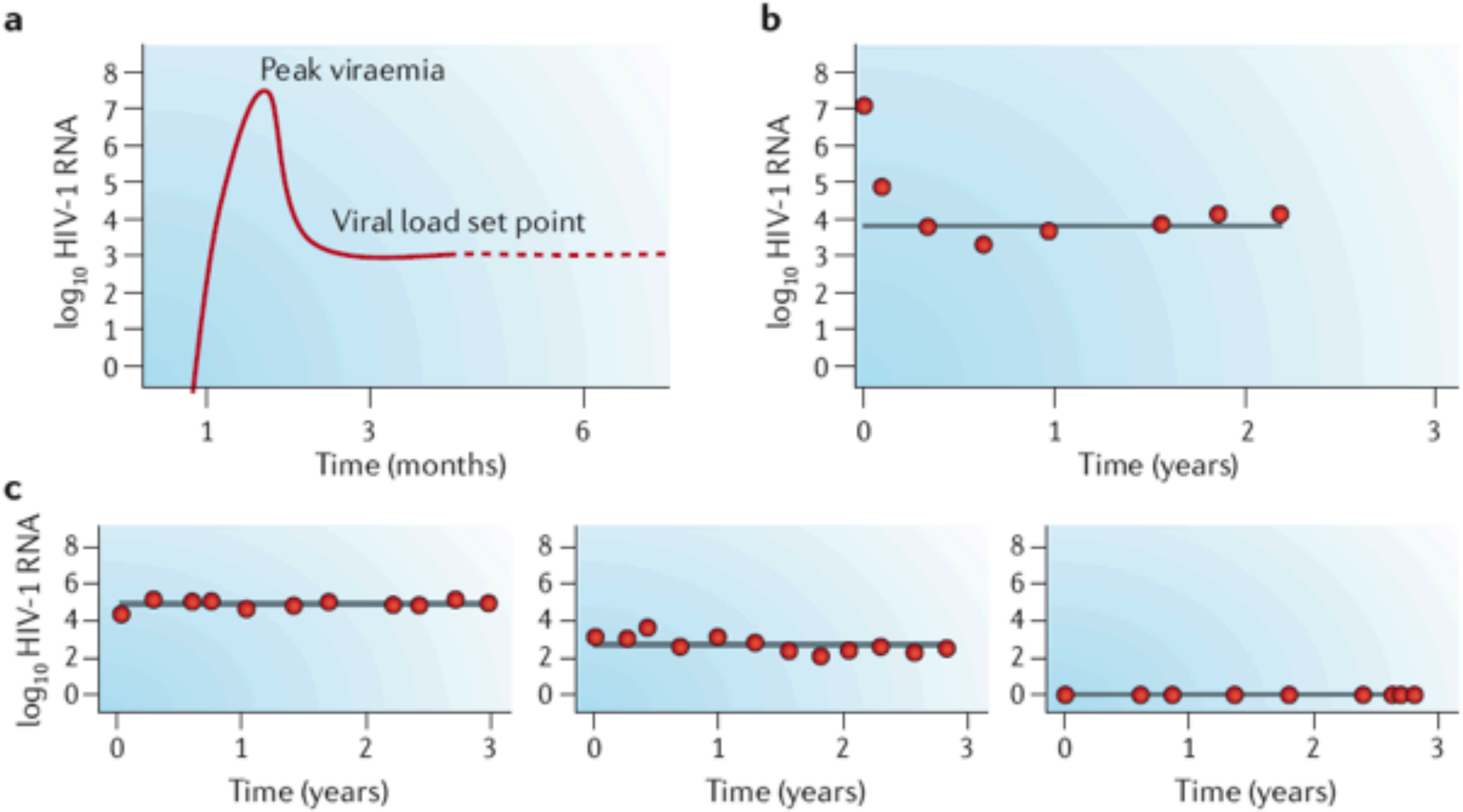
Exposure

Infection

Genome-wide research demands

- precise phenotype
- careful selection of patients
- efficient genotyping
- powerful analysis

Phenotype: HIV viral load at set point

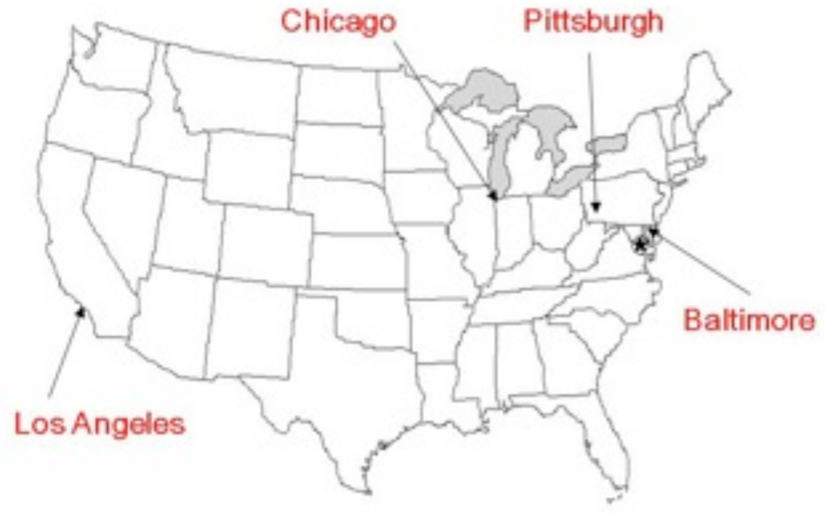


Telenti & Goldstein, Nat Rev Microbiol 2006

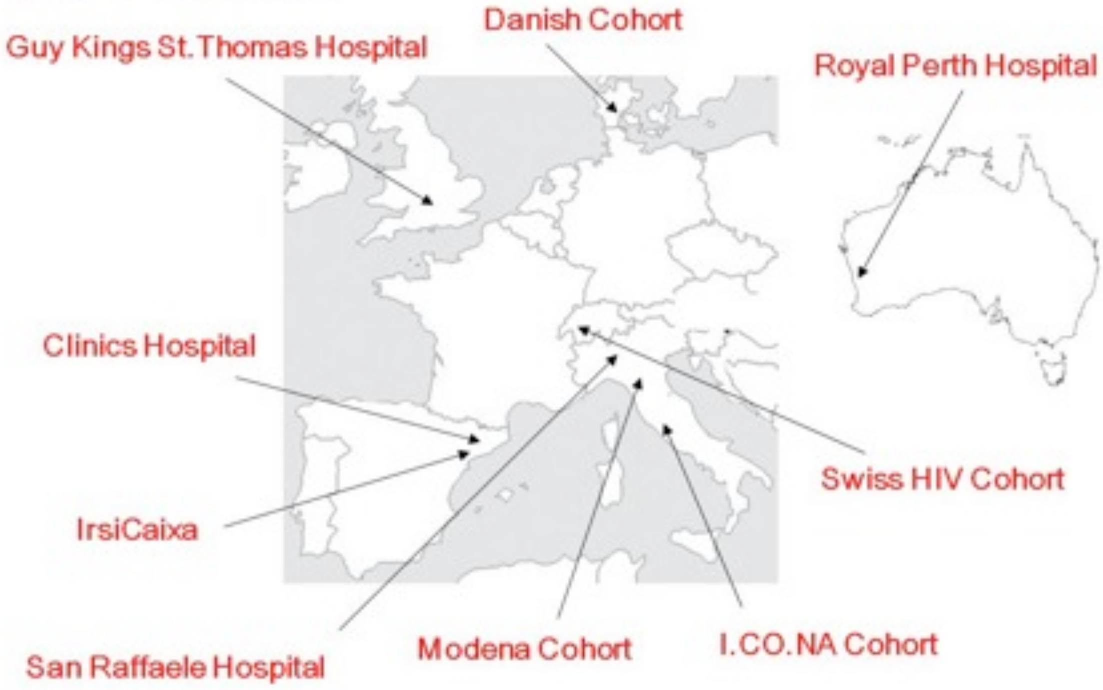
Patients / Cohorts

- ~2500 white patients
- High quality viremia data
- Genetic consent

MACS



Euro-CHAVI



Genotyping

WG chips:

500K to >1 mio single nucleotide polymorphisms (SNPs)

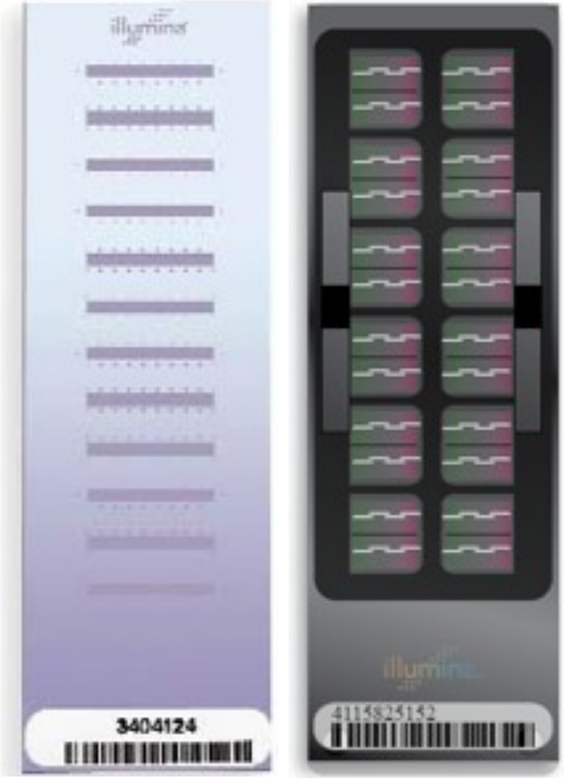
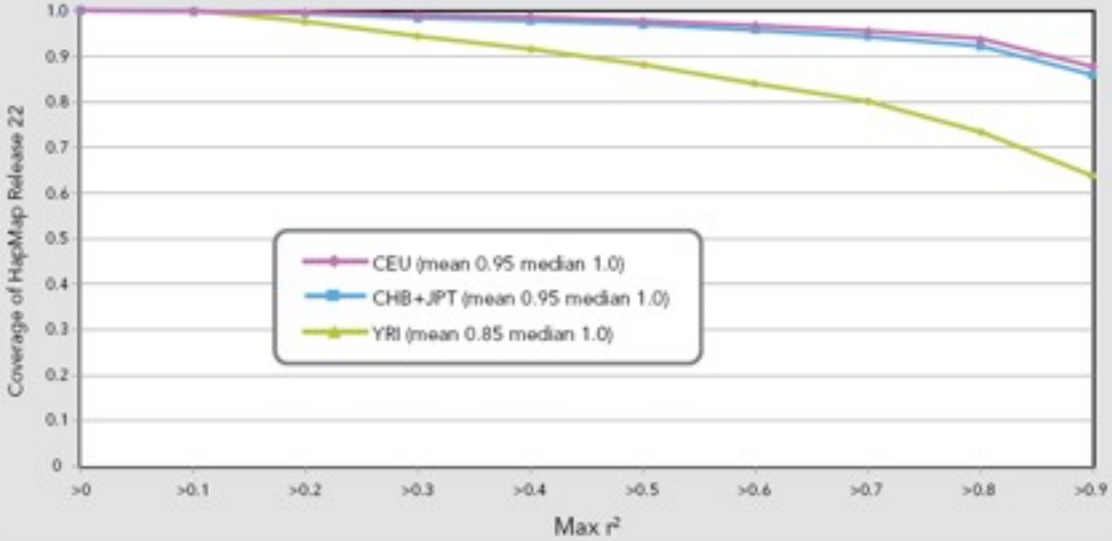


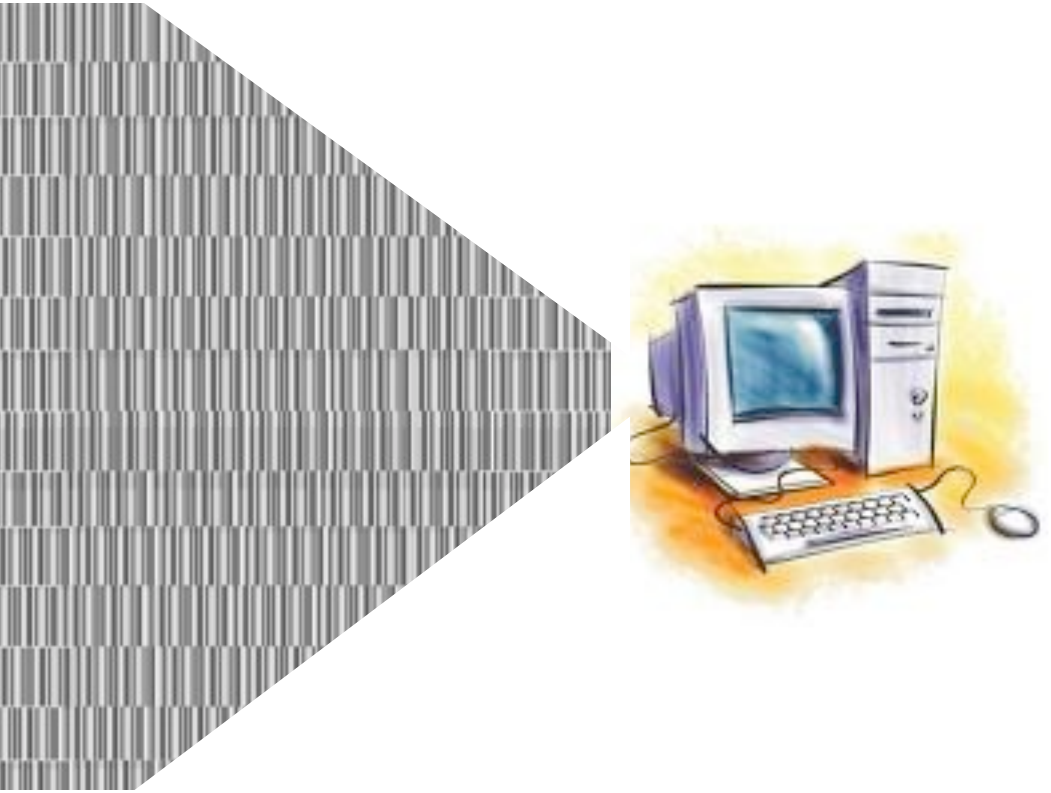
FIGURE 5: HUMAN1M-DUO GENOMIC COVERAGE



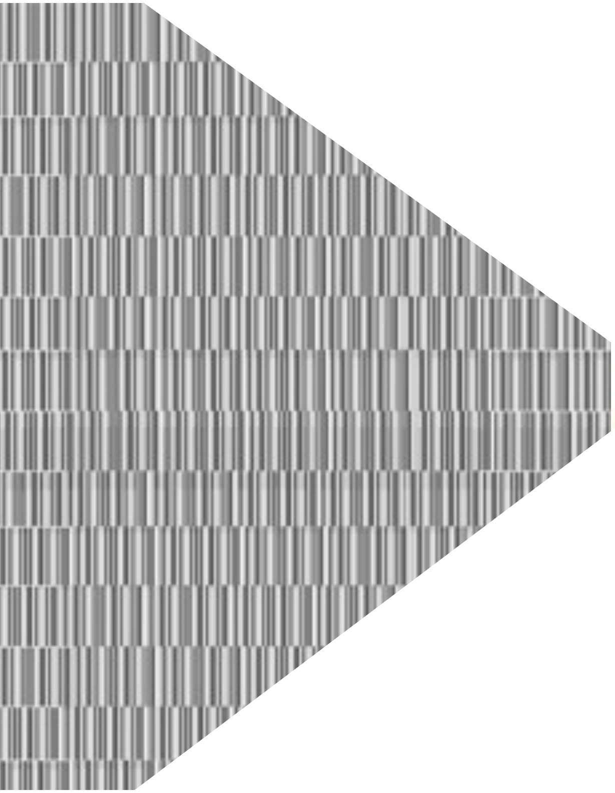
The Human1M-Duo BeadChip content covers the majority of HapMap common variation in three distinct populations. Graphs are based on the HapMap release 23 data set of > 2.3 million common SNPs.

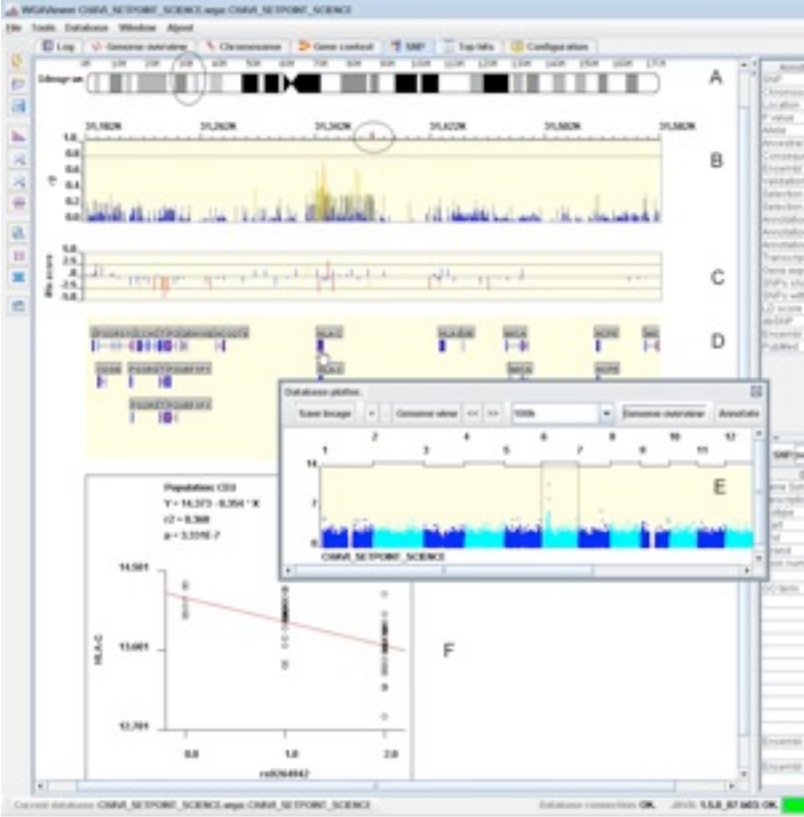
Analysis

Analysis

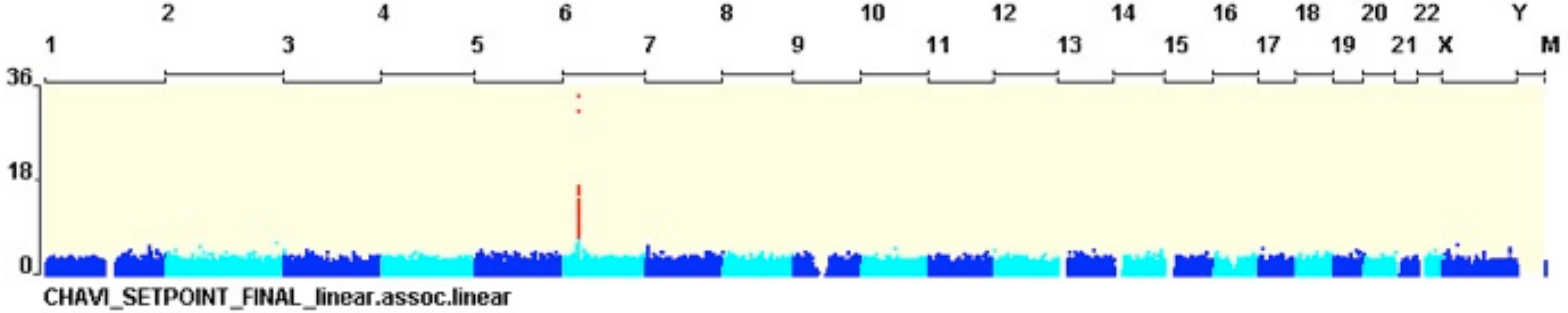


Analysis

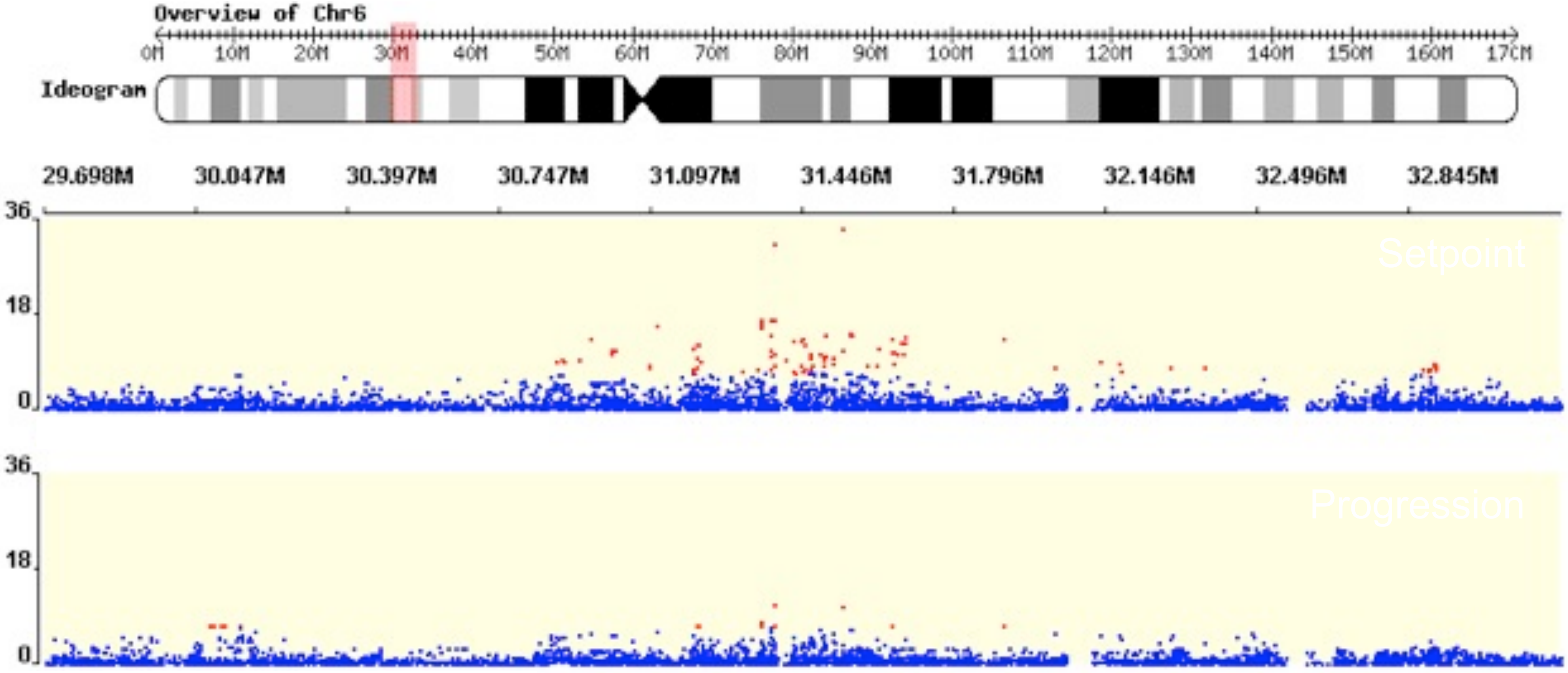




Genetics of HIV-1 control: results

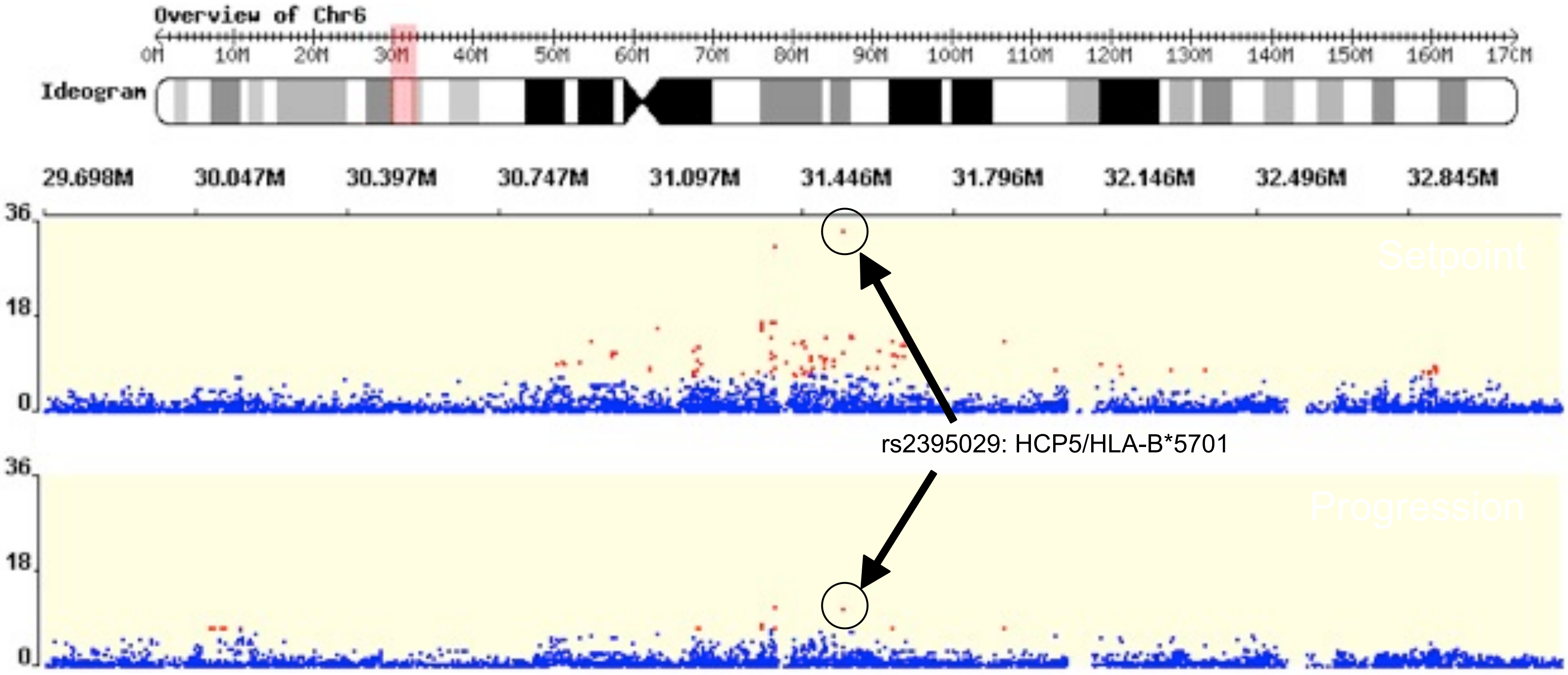


Genetics of HIV-1 control: results



Fellay et al. Science 2007

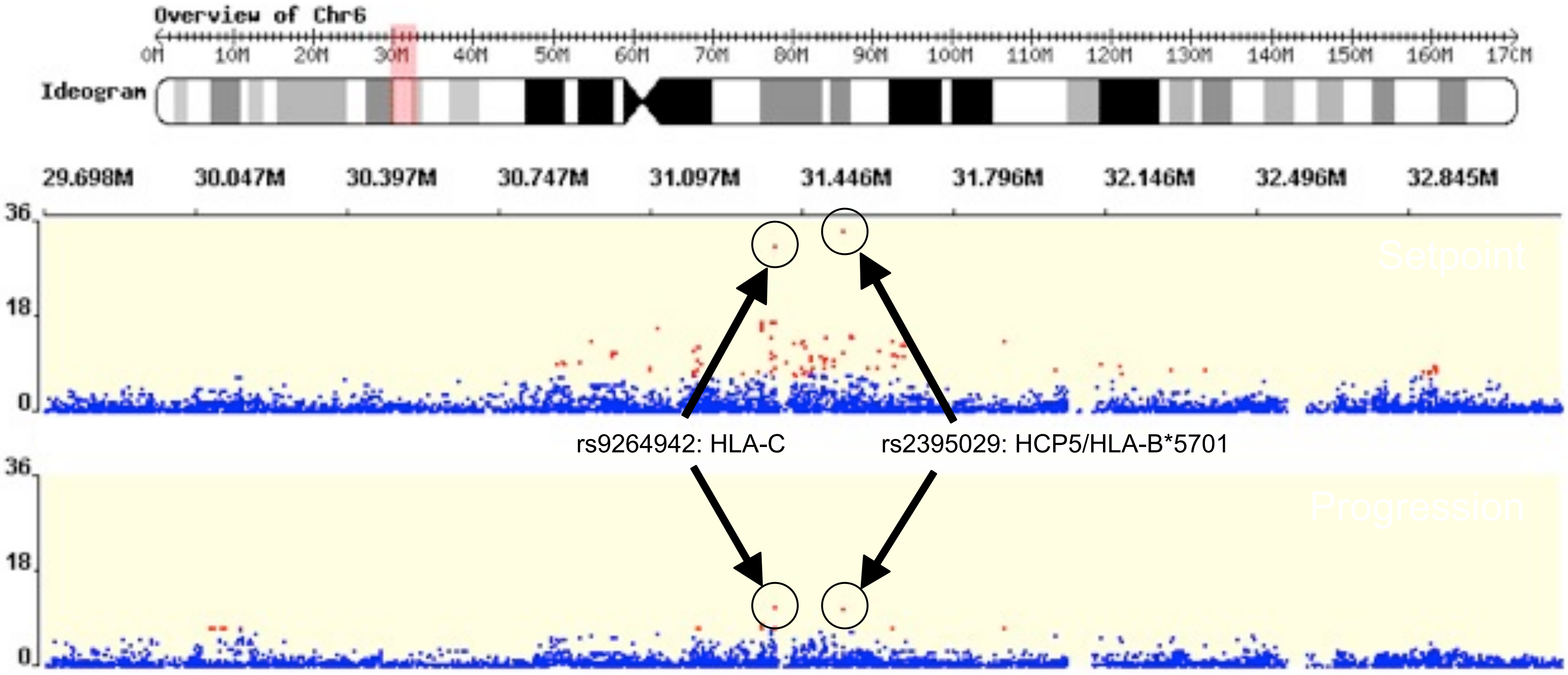
Genetics of HIV-1 control: results



rs2395029: HCP5/HLA-B*5701

Fellay et al. Science 2007

Genetics of HIV-1 control: results

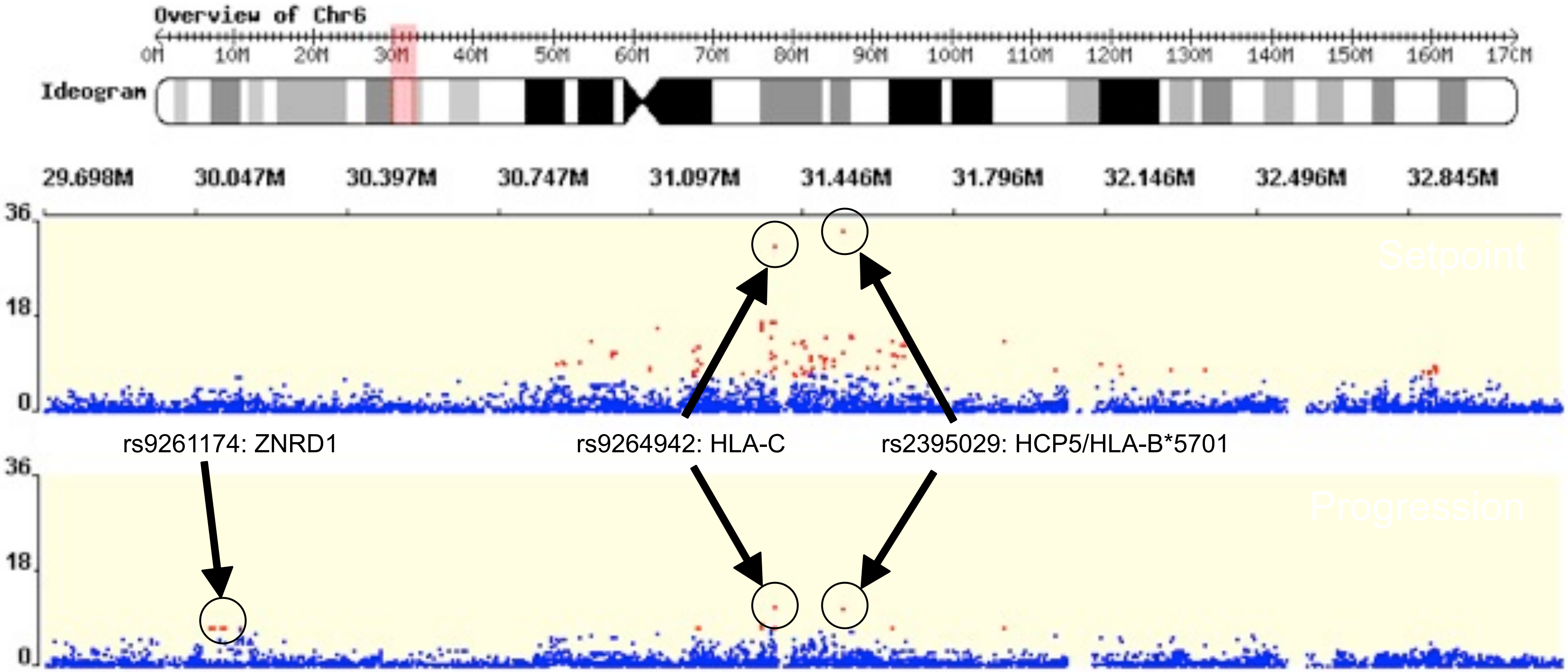


rs9264942: HLA-C

rs2395029: HCP5/HLA-B*5701

Fellay et al. Science 2007

Genetics of HIV-1 control: results



rs9261174: ZNRD1

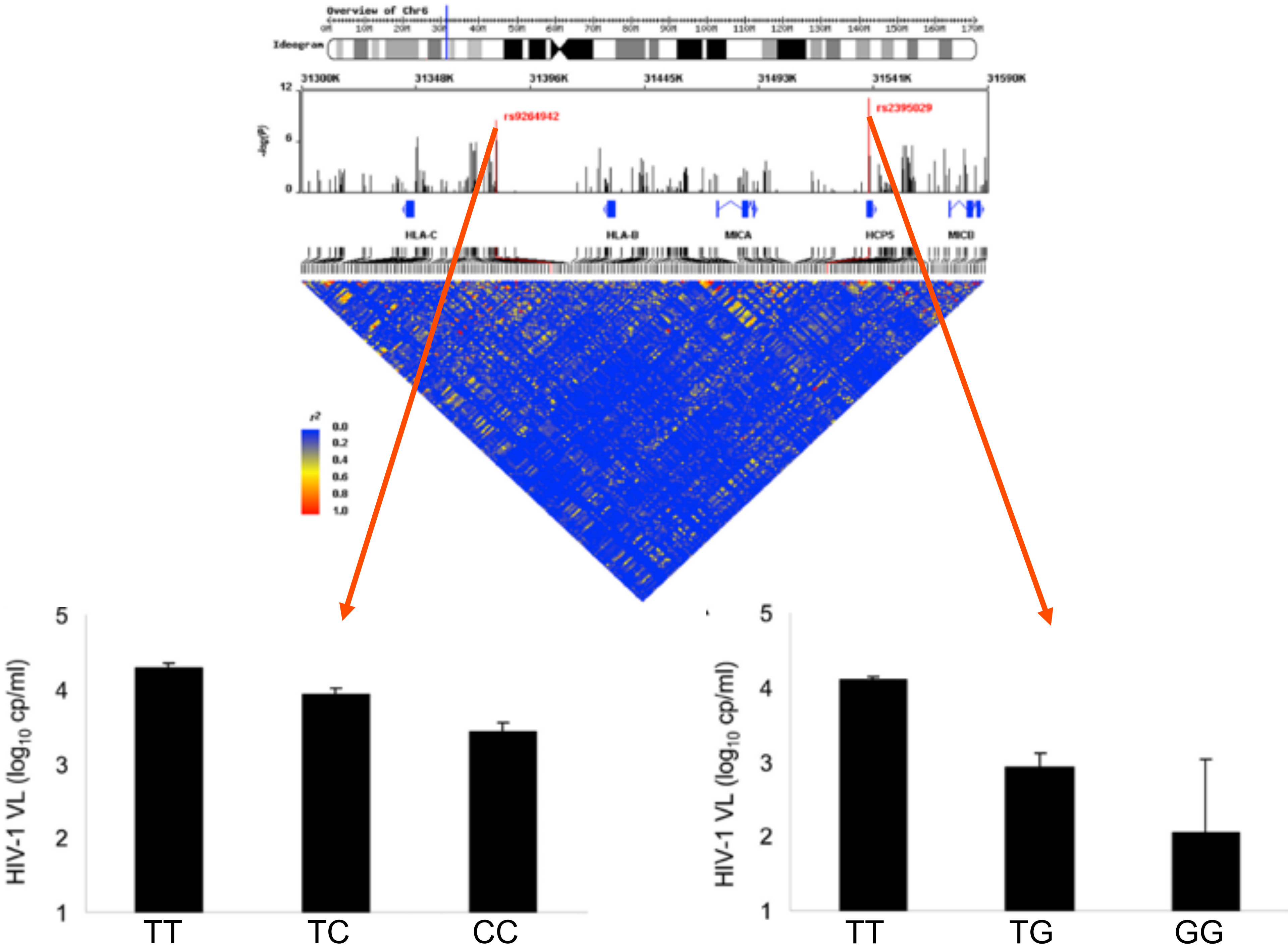
rs9264942: HLA-C

rs2395029: HCP5/HLA-B*5701

Fellay et al. Science 2007

Gene & SNP	P-value for association with HIV-1 viral load at setpoint N=2362	P-value for association with protection against progression (CD4 <350) N=1071
HCP5 / HLA-B*5701 rs2395029	4.5E-35	1.2E-11
HLA-C rs9264942	5.9E-32	7.4E-12
ZNRD1 / RNF39 rs9261174	1.1E-04	3.8E-08
CCR5 Δ32 het rs333	1.7E-10	2.6E-06

Bonferroni threshold for genome-wide significance: 5E-08

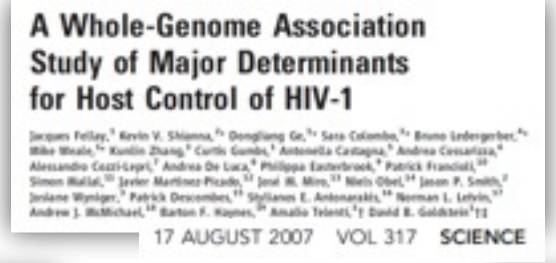

Overview of Chr6
0M
10M
20M
30M
40M
50M
60M
70M
80M
90M
100M
110M
120M
130M
140M
150M
160M
170M
Ideogram
31300K
31348K
31396K
31445K
31493K
31541K
31590K
12
6
0
-log(P)
rs9264942
rs2395029
HLA-C
HLA-B
MICA
HCP5
MICB
r^2
0.0
0.2
0.4
0.6
0.8
1.0
HIV-1 VL ($\log_{10}$ cp/ml)
5
4
3
2
1
TT
TC
CC
HIV-1 VL ($\log_{10}$ cp/ml)
5
4
3
2
1
TT
TG
GG

Genes and viral load to predict disease progression

(no progression after 5 years, %)

Sequencing example

Host genetics of HIV-1 control



A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1

Jacques Fellay, Kevin V. Shianna, Dongliang Ge, Sara Colombo, Bruno Ledergerber, Mike Weale, Kunlin Zhang, Curtis Gumbs, Antonella Castagna, Andrea Cossarizza, Alessandro Cozzi-Lepri, Andrea De Luca, Philippa Easterbrook, Patrick Francioli, Simon Mallal, Javier Martinez-Picado, José M. Miro, Niels Obel, Jason P. Smith, Josiane Wyniger, Patrick Descombes, Stylianos E. Antonarakis, Norman L. Letvin, Andrew J. McMichael, Barton F. Haynes, Amalio Telenti, David B. Goldstein

17 AUGUST 2007 VOL 317 SCIENCE

JID 2010:201 (15 April)

MAJOR ARTICLE

Host Determinants of HIV-1 Control in African Americans

Kimberly Pelak, David B. Goldstein, Nicole M. Walley, Jacques Fellay, Dongliang Ge, Kevin V. Shianna, Curtis Gumbs, Xiaojiang Gao, Jessica M. Maia, Kenneth D. Cronin, Shehnaz K. Hussain, Mary Carrington, Nelson L. Michael, Amy C. Weintrob, and the Infectious Disease Clinical Research Program HIV Working Group, on behalf of the National Institute of Allergy and Infectious Diseases Center for HIV/AIDS Vaccine Immunology (CHAVI)

GWAS

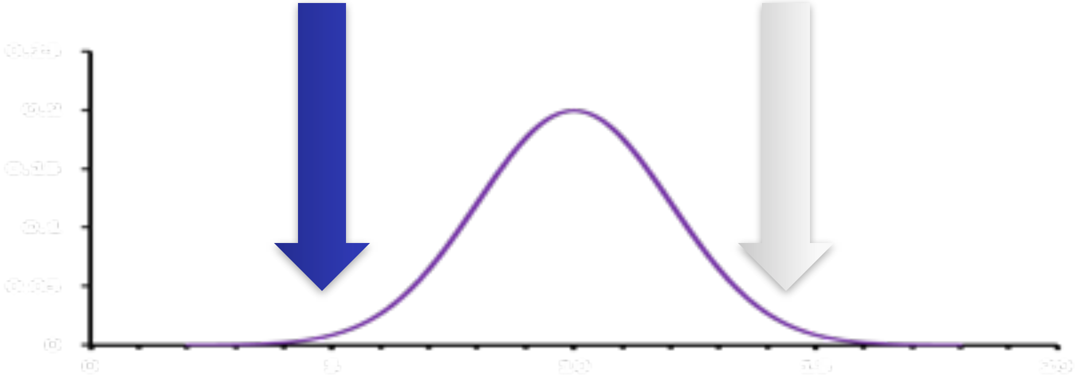
OPEN ACCESS Freely available online

PLoS GENETICS

Common Genetic Variation and the Control of HIV-1 in Humans

Jacques Fellay, Dongliang Ge, Kevin V. Shianna, Sara Colombo, Bruno Ledergerber, Elizabeth T. Cirulli, Thomas J. Urban, Kunlin Zhang, Curtis E. Gumbs, Jason P. Smith, Antonella Castagna, Alessandro Cozzi-Lepri, Andrea De Luca, Philippa Easterbrook, Huldrych F. Günthard, Simon Mallal, Cristina Mussini, Judith Dalmau, Javier Martinez-Picado, José M. Miro, Niels Obel, Steven M. Wolinsky, Jeremy J. Martinson, Roger Detels, Joseph B. Margolick, Lisa P. Jacobson, Patrick Descombes, Stylianos E. Antonarakis, Jacques S. Beckmann, Stephen J. O'Brien, Norman L. Letvin, Andrew J. McMichael, Barton F. Haynes, Mary Carrington, Sheng Feng, Amalio Telenti, David B. Goldstein, NIAID Center for HIV/AIDS Vaccine Immunology (CHAVI)

Next-generation sequencing of extreme phenotypes



Sequencing of extreme HIV progressors

1. Rapid Progressors
   - Known date of seroconversion
   - CD4 <350 in less than 3 years
   - [Severe PHI = immediate CD4 depletion without spontaneous recovery]

2. Controllers
   - VL >50 cp/ml
   - Excluding HLA-B*57, B*27 and B*5801

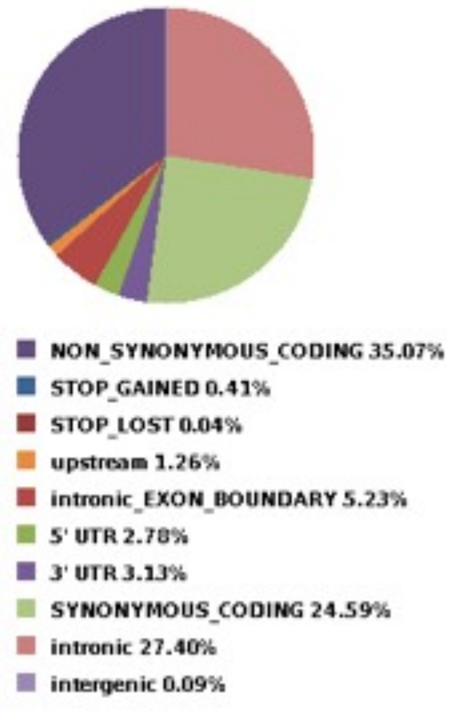
Exome sequencing in 31 rapid progressors and 10 controllers

Overview of genetic variation

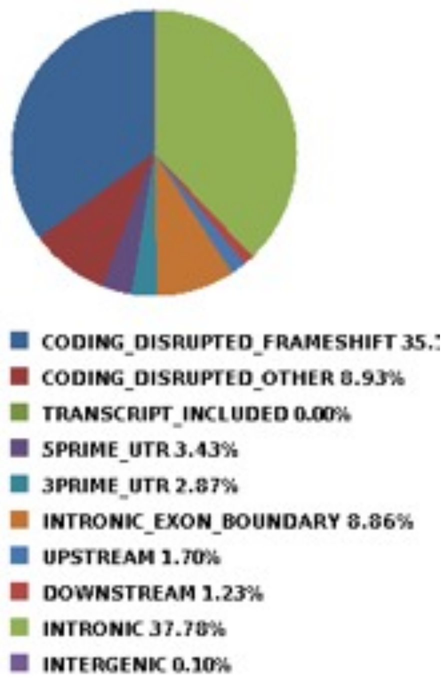
- Mean coverage: 72x

- Total SNVs: 101057
- Novel SNVs: 40385 (40%)
- Total indels: 12149
- Novel indels: 5330 (45.5%)

SNVs



indels



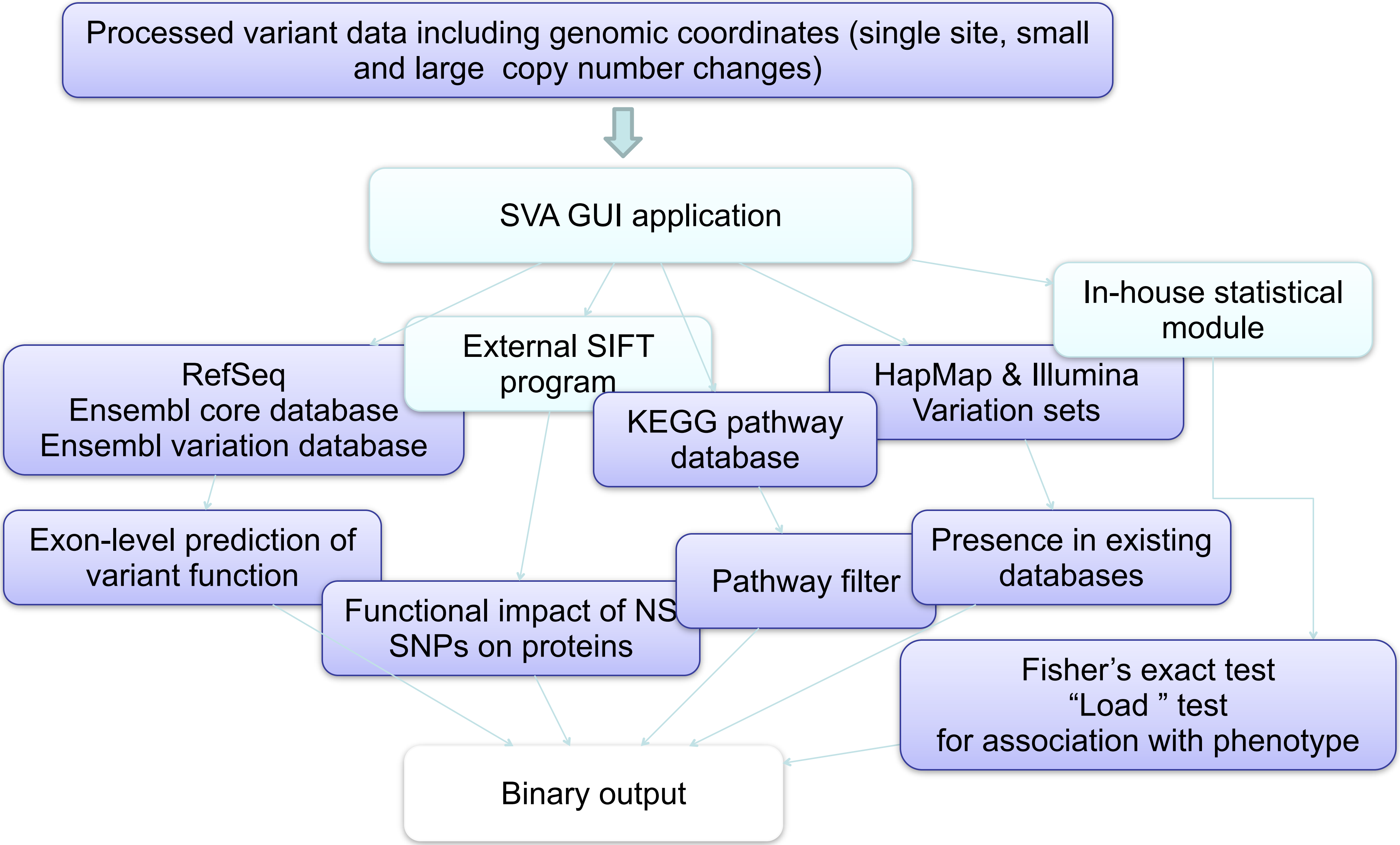
Analysis of genetic variants

- Single variant analysis (ATAV)
  - Case-control comparison of single variants (SNVs and indels) using Fisher’s exact tests for allelic, dominant, recessive, and genotypic models, plus Cochran-Armitage trend test

- Ranking of putatively functional variants (SVA)
  - listing of homozygous or heterozygous variants observed mostly (or only) in cases, ranked by numbers

- Gene prioritization (SVA and ATAV)
  - Case-control comparison of genes carrying key functional variants, using Fisher’s exact tests with assessment of genome-wide significance by permutations

How does it work?



Processed variant data including genomic coordinates (single site, small and large copy number changes)

SVA GUI application

In-house statistical module

External SIFT program

RefSeq
Ensembl core database
Ensembl variation database

HapMap & Illumina
Variation sets

KEGG pathway database

Exon-level prediction of variant function

Functional impact of NS SNPs on proteins

Pathway filter

Presence in existing databases

Fisher's exact test
"Load " test
for association with phenotype

Binary output

Single variant analysis

variant	RS	gene	function	RP	VC	Ctrls	P_value
19_59711073_T	-	**LAIR2 / CD306**	STOP_GAINED	1/4/26	0/0/10	0/19/208	0.05 (genotypic:RPvsCtrls)
5_86731030_G	rs2230641	**CCNH**	NS	2/12/17	0/0/10	10/59/160	0.009 (allelic:RPvsVC)
16_55617854_C	rs28438857	**NLRC5**	NS	3/7/21	0/1/9	0/59/167	0.002(recessive:RPvsCtrls)
11_60533649_A	rs12360861	**CD6**	NS	3/8/19	0/3/7	2/48/160	0.02 (trend:31vsCtrls)
4_74921673_INS_T	-	**CXCL6**	FRAME SHIFT	0/2/29	0/0/10	0/2/226	0.018 (trend:RPvsCtrls)
1_26517124_A	-	**CD52**	NS	0/3/28	0/0/10	0/0/229	0.0006 (trend:RPvsCtrls)
1_12108645_G	rs2230625	**TNFRSF8 / CD30**	NS	0/3/28	0/3/6	0/5/210	0.0009 (trend:VCvsCtrls)
1_158052037_T	rs61823162	**FCRL6**	STOP_GAINED	3/3/25	0/1/9	7/52/170	0.06 (genotypic:RPvsCtrls)

Next steps

1. **More “extreme” samples:**

   ➔ MACS: 25 rapid progressors
   25 controllers

2. **More sequence:**

   ➔ Whole genome sequencing

From GWAS to sequencing, and beyond…

- Only a limited amount of the genetic basis for much phenotypic variation has been located by GWAS
  - ➢ Still much ‘missing heritability’

- Rare and/or causal variants will be identified by sequencing

- Data integration and systems approaches represent the next frontier