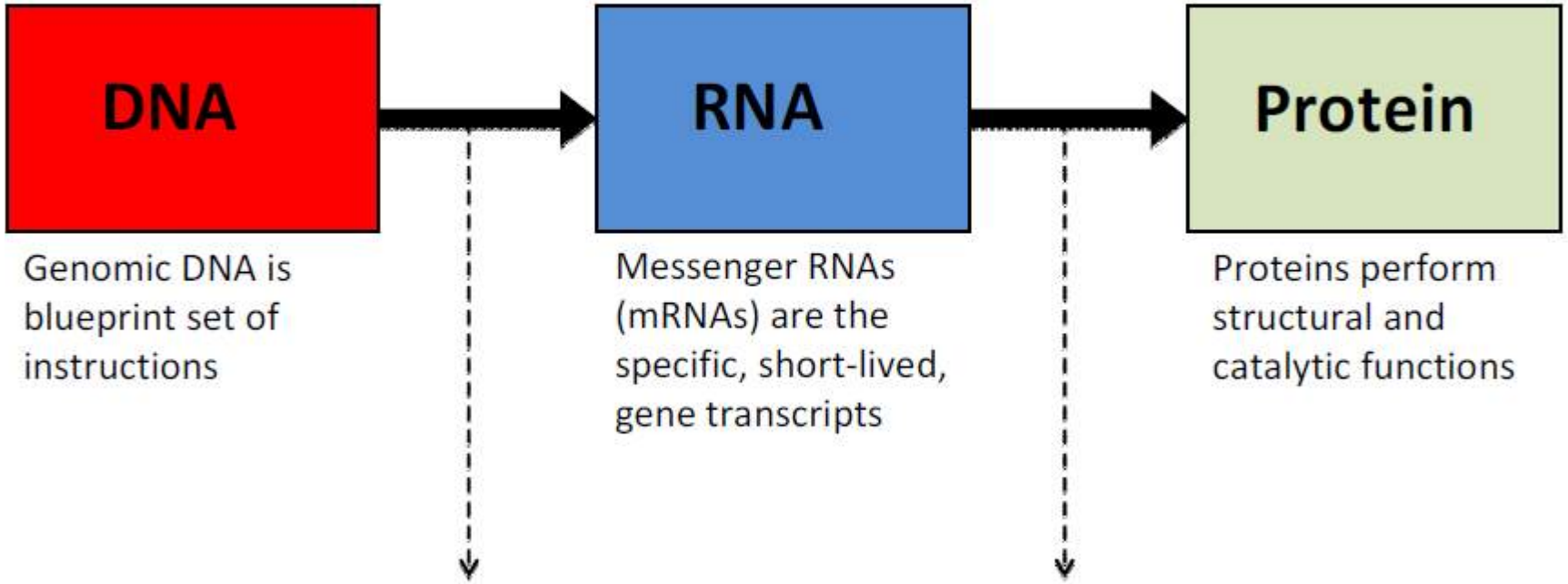
Meredith Yeager, Ph.D.
Scientific Director,
Cancer Genomics Research Laboratory

Genetic and Genomics Laboratory Tools and Approaches

Radiation Epidemiology & Dosimetry Course

Central Dogma of Biology:



Sequencing of the Human Genome

Nature **409**, 860-921 (2001)

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

Science **291**, 1304-1351 (2001)

THE HUMAN GENOME

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]
Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1]
Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1]

The Human Genome – 2015

- ~3.1 billion bases (A, C, G, T)
- ~20 thousand protein-coding genes, many non-coding RNAs
- Annotation ongoing – the initial sequencing in 2001 is still being refined, assembled and annotated, even now
- Variation (polymorphism) present within humans
  - Population-specific
  - Cosmopolitan

Types of polymorphisms

- Single nucleotide polymorphisms (SNPs)
  - Common SNPs are defined as > 5% in at least one population
  - Abundant in genome (~50 million and counting)
- Repeats of DNA (long, short, complex, simple), insertions/deletions
- **A small fraction of SNPs and other types of variation** are very or slightly deleterious and may contribute by themselves or with other genetic or environmental factors to a phenotype or disease

ATGGAACGA(G/C)AGGATA(T/A)TACGCACTATGAAG(C/A)CGGTGAGAGG

SNPs & function

- Vast majority are “silent”
  - No known functional change
- Alter function of gene product
  - Change sequence of protein
- Alter gene expression/regulation
  - Promoter/enhancer
  - mRNA stability
  - Small RNA binding sites
  - Disrupt CpG site

International HapMap Project

Home | About the Project | Data | Publications | Tutorial

中文 | English | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "**About the International HapMap Project**" for more information.

Project Information

About the Project
HapMap Publications
HapMap Tutorial
HapMap Mailing List
HapMap Project Participants

Project Data

HapMap Genome Browser release #28 (Phases 1, 2 & 3 - merged genotypes & frequencies)
HapMap3 Genome Browser release #3 (Phase 3 - genotypes & frequencies)
HapMap Genome Browser release #27 (Phase 1, 2 & 3 - merged genotypes & frequencies)
HapMap3 Genome Browser release #2 (Phase 3 - genotypes, frequencies & LD)
HapMap Genome Browser release#24 (Phase 1 & 2 - full dataset)
GWAs Karyogram
HapMart
HapMap FTP
Bulk Data Download
Data Freezes for Publication
ENCODE Project
Guidelines For Data Use

Useful Links

TSC SNP Downloads
HapMap Samples at Coriell Institute
HapMap Project Press Release
NHGRI HapMap Page
NCBI Variation Database (dbSNP)
Japanese SNP Database (JSNP)

News

- 2011-06-13: **HapMap help desk announcement**

  There was a problem with the HapMap help desk system. In the past several weeks, emails sent to hapmap-help@ncbi.nlm.nih.gov did not reach the help desk, and thus user requests were not addressed. Please resend your email request if you sent emails to the HapMap help desk in the past several weeks. Sorry for the inconvenience.

- 2011-04-20: **Hapmap help desk service interruption notice**

  There will be no help desk support from 05/03/2011 to 05/23/2011. Sorry for the inconvenience.

- 2011-02-02: **Haploview issues with rel 28 data**

  Recently, there are several questions about Haploview data format errors when users tried to analyze HapMap release 28 data. The current Haploview version (4.2) does not recognize the new individuals in release 28 and the software will generate an error similar to "Hapmap data format error: NA18876" when trying to open the data.

  Haploview is developed and maintained by an organization different from HapMap. Please contact Haploview help desk (haploview@broadinstitute.org) for questions specific to this software.

- 2011-01-19: **HapMap phase II recombination rate on GRCh37**

  The liftover of the

- 2010-08-18: **Hap**

  Genotypes and f
  bulk download

- 2010-05-28: **Hap**

  Genotypes and f
  and also **availab**

- 2010-05-28: **Hap**

  Copy Number Va

- 2009-12-10: **Cor**

  Phased haplotyp
  inconvenience th

- 2009-12-02: **Hap**

1000 Genomes

A Deep Catalog of Human Genetic Variation

Home | About | Data | Analysis | Participants | Contact | Browser | Wiki | FTP search | Search

LATEST ANNOUNCEMENTS

MONDAY JULY 02, 2012

Phase 1 analysis results including chrY and chrMT variant calls.

Analysis results based on our phase1 integrated variant call set are now available.

This includes chrY and chrMT variant calls, functional annotation of our variant calls and local area ancestry inference for our admixed populations.

Full details of the directory contents can be found on this webpage.

Data Access links: EBI / NCBI

README: EBI / NCBI

Recent project announcements

THURSDAY SEPTEMBER 06, 2012

Genome Accessibility information now available on the 1000 Genomes Browser

Two Accessibility Tracks have now been added to the 1000 Genomes Browser

This information was built using sequence data from the phase1 dataset

The two tracks are called the 1000 Genomes Pilot Accessibility Mask and the 1000 Genomes Strict Accessibility Mask.

There is a README which describes how this data set was created. The raw bed and fasta files are also available in the accessible genome ftp directory

TUESDAY MAY 22, 2012

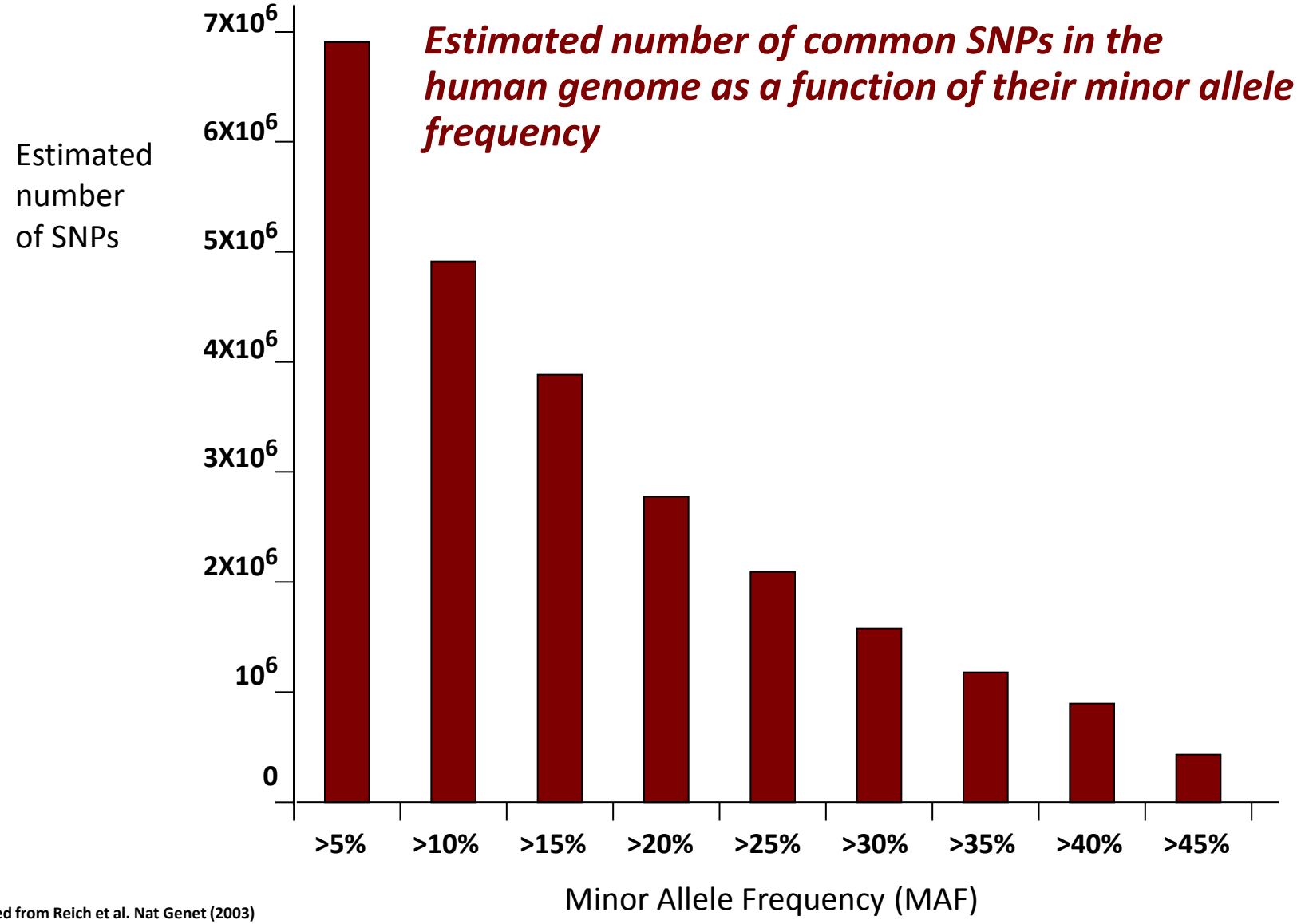
New Sequence Data is Available

NAVIGATION

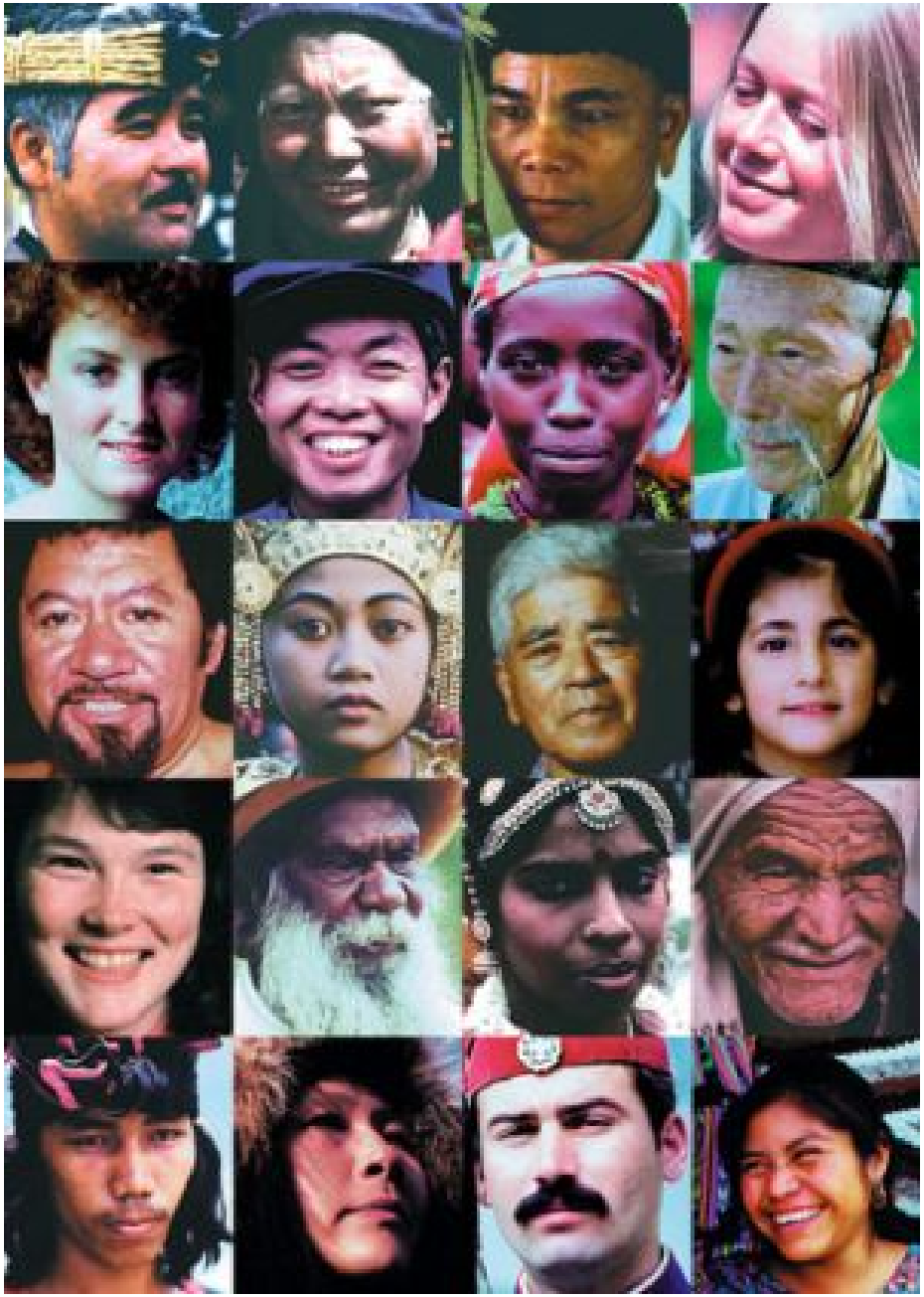
- Frequently Asked Questions

LINKS

All Project Announcements
Sample and Project Information
Media Archive
Download the 1000 Genomes Pilot Paper
Project Contacts
RSS Feed
Twitter

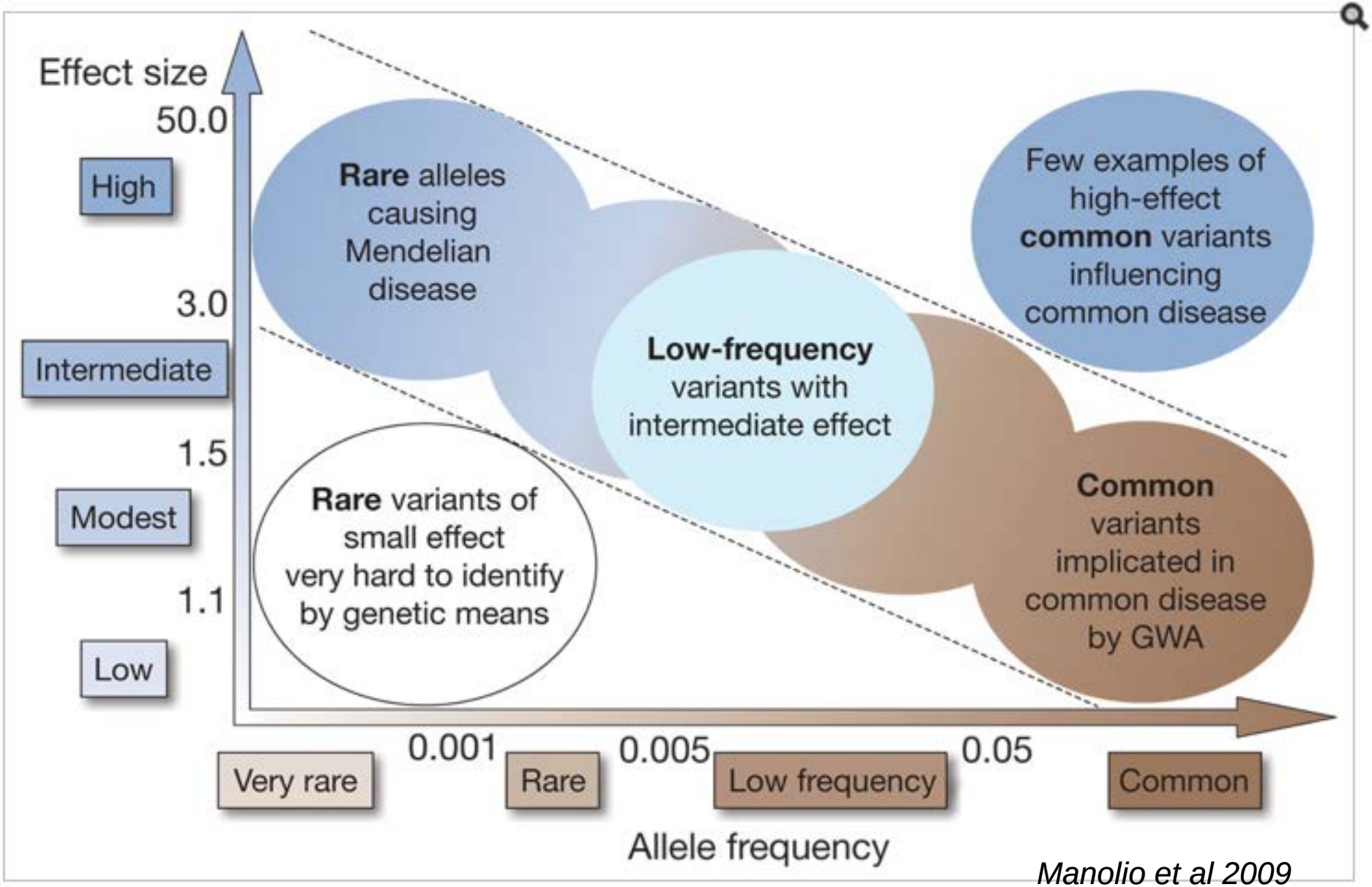


Adapted from Reich et al. Nat Genet (2003)



Cancer Genomics Research Laboratory

Mapping Genetic Susceptibility

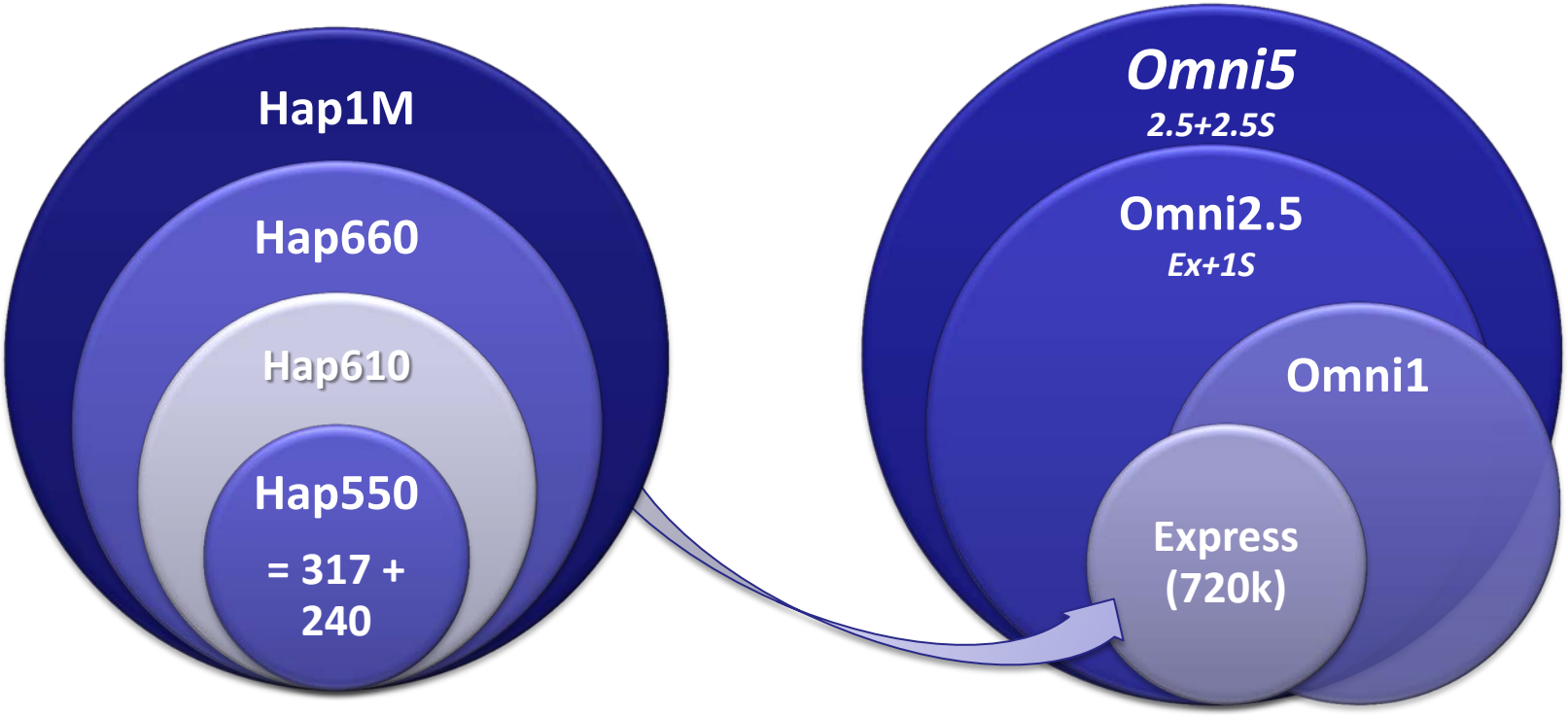


Manolio et al 2009

SNP genotyping

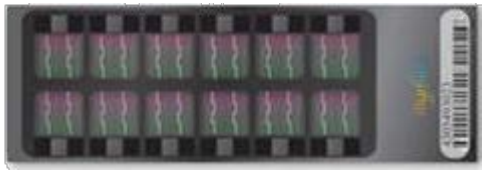

HiScan
illumina

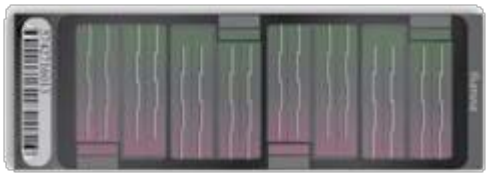
Illumina "whole-genome" SNP microarrays



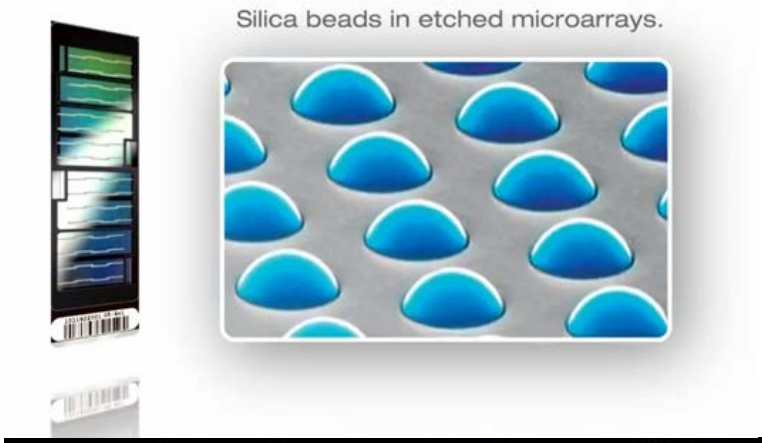
HUMAN HAP (HAPMAP)

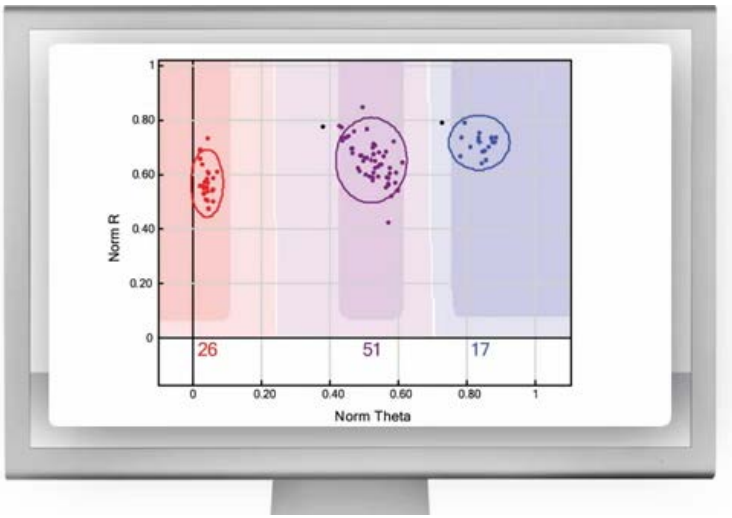
HUMAN OMNI (+1K GENOMES)

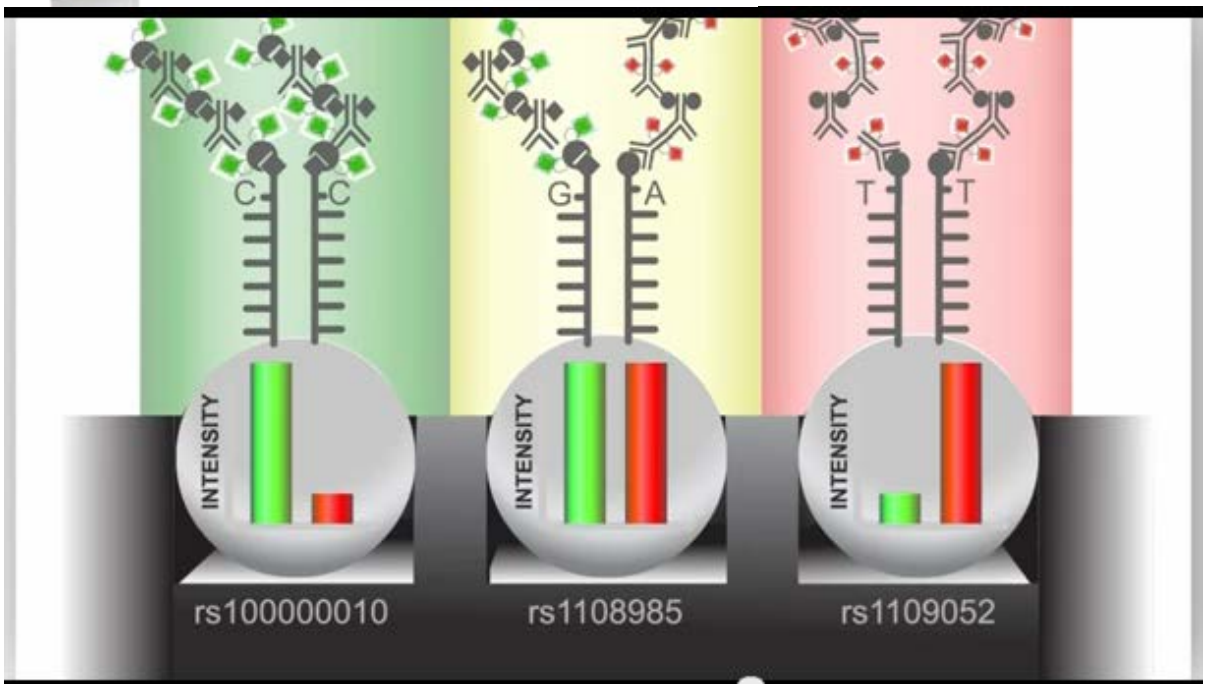




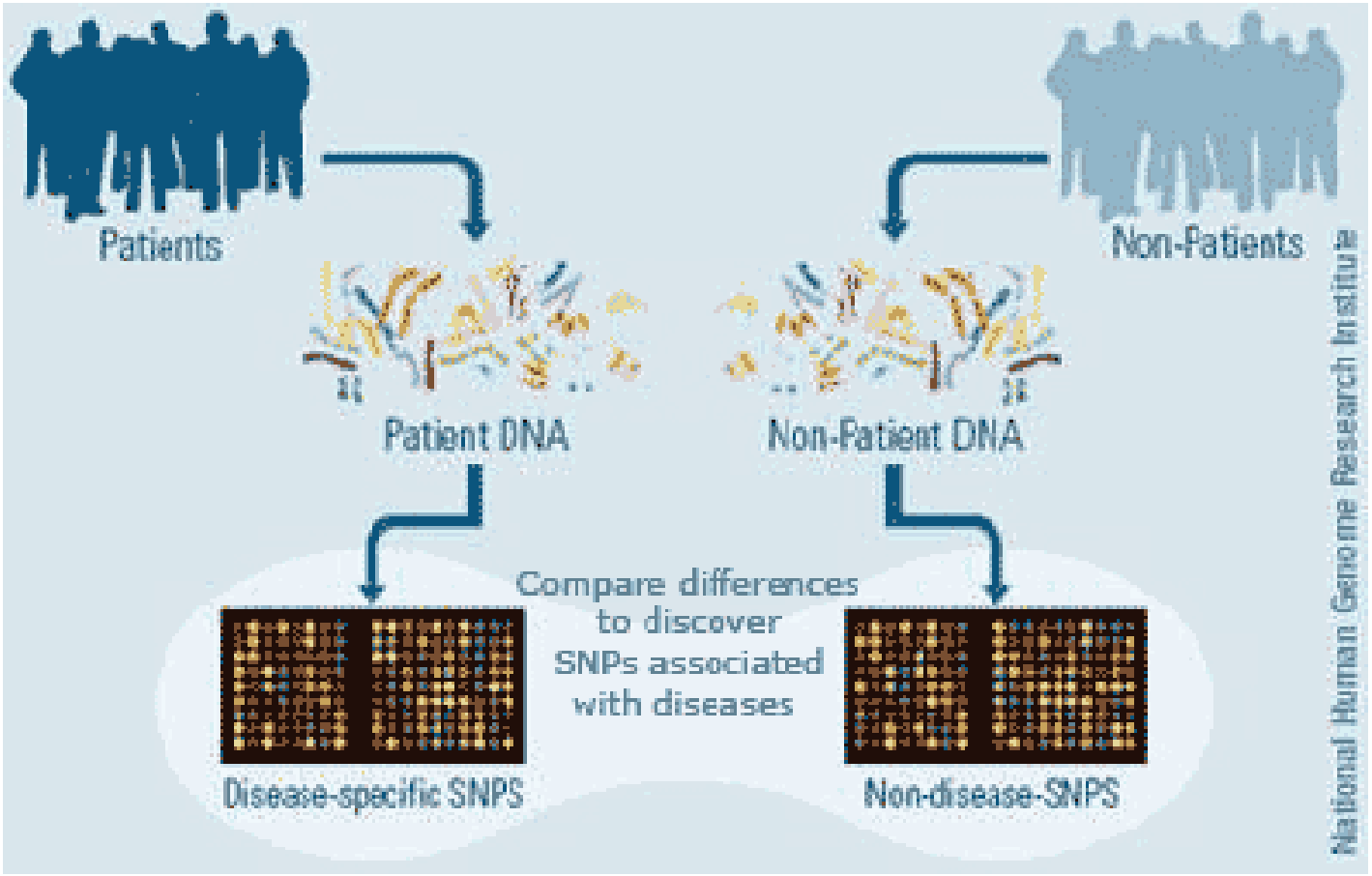
Infinium Technology: Two color allele-specific detection at each locus

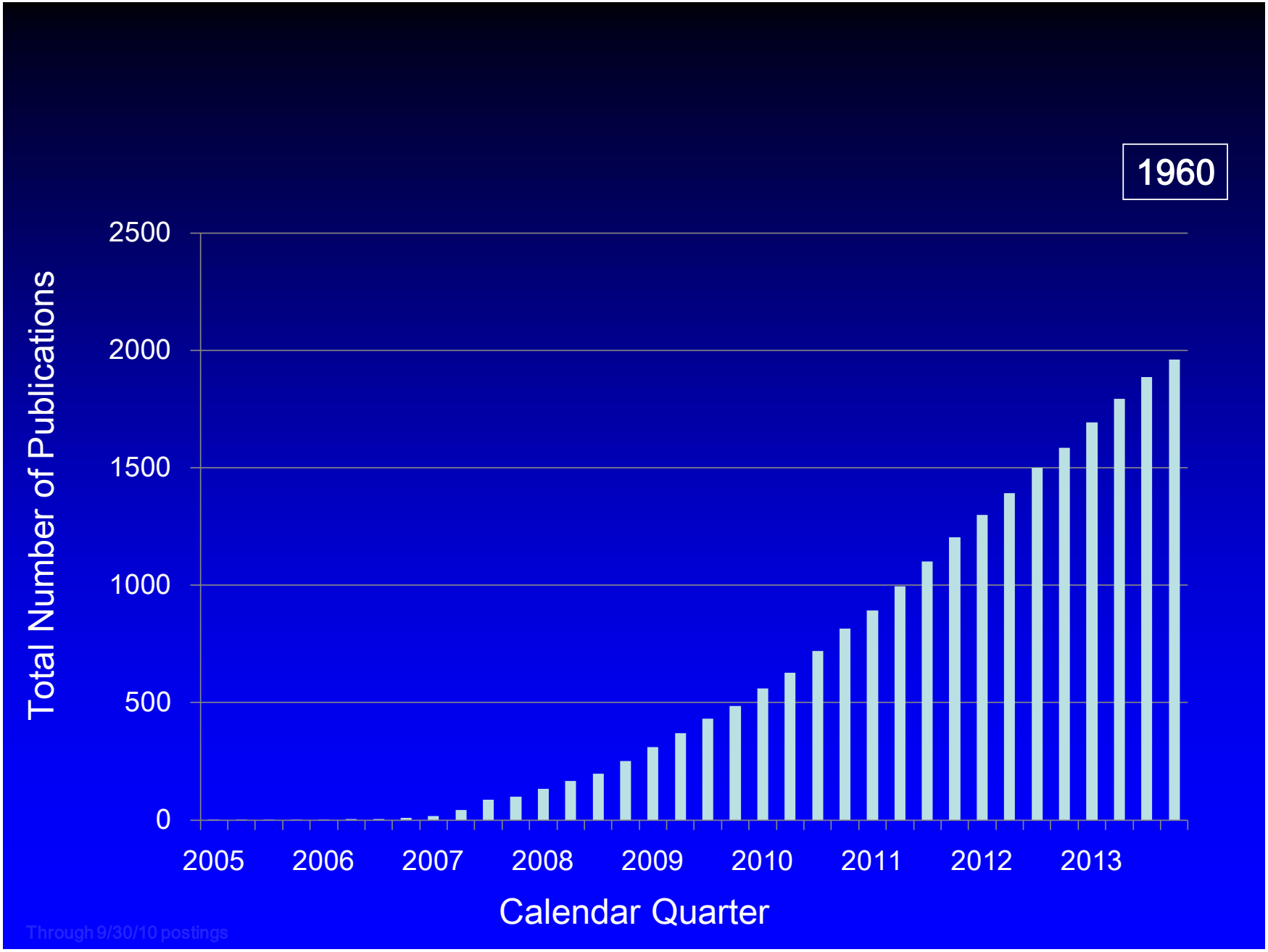






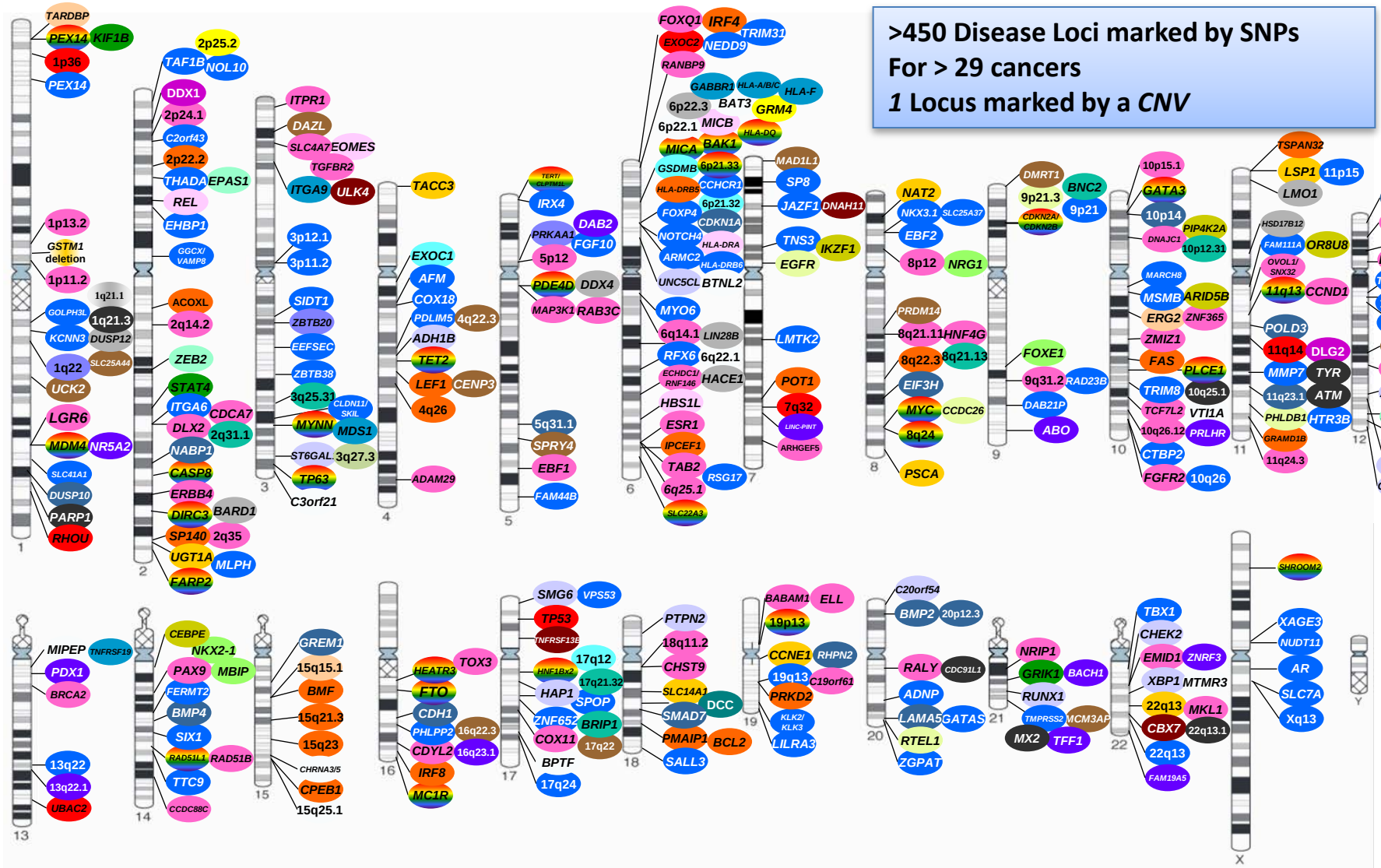
Basic principle of genetic association studies in unrelated individuals





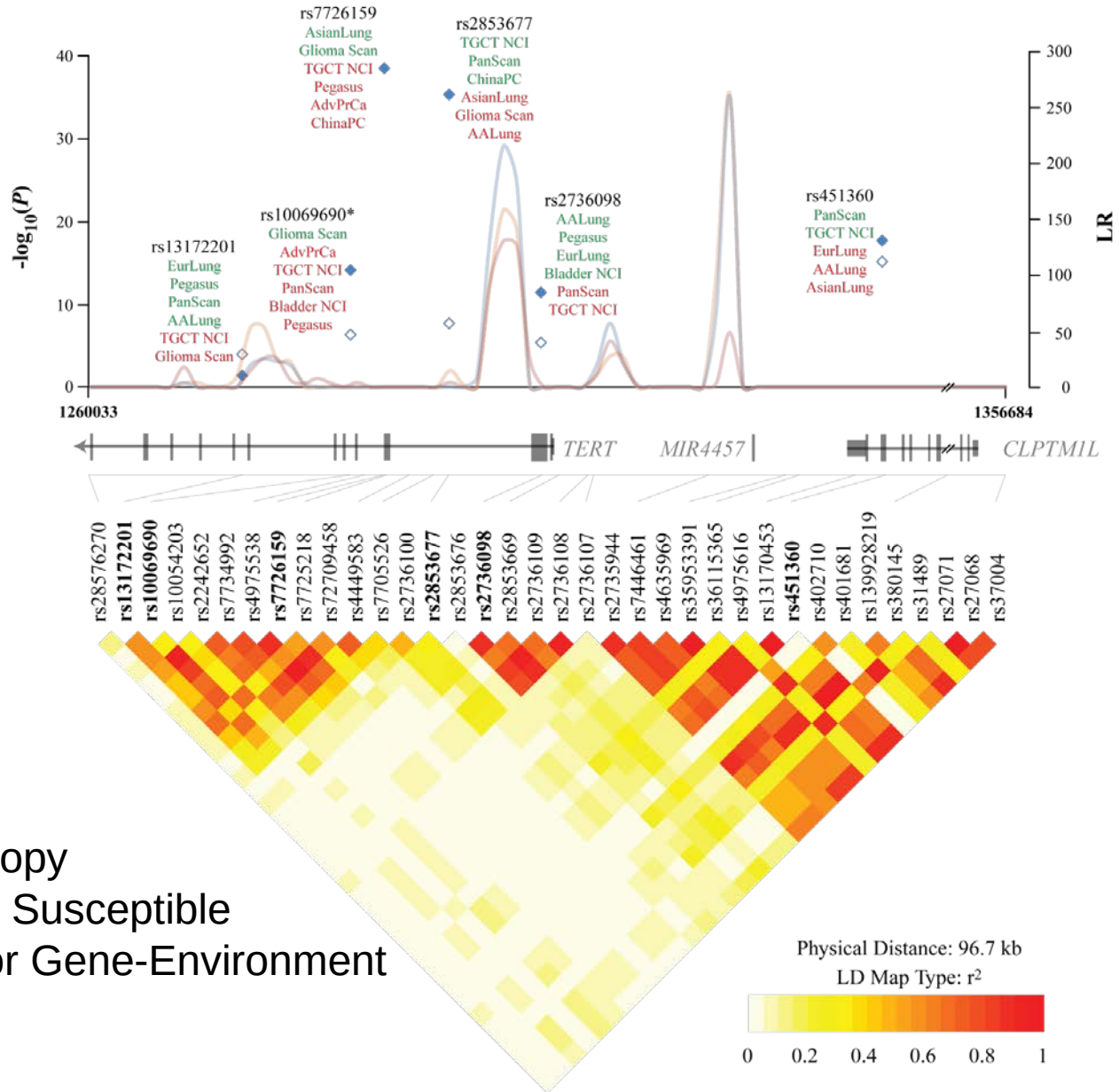
https://www.genome.gov/26525384

Published Cancer GWAS Etiology Hits: Sep 2014



11 Basal Cell | 13 Bladder | 82 Breast | 7 Cervical | 31 CLL | 26 Colorectal | 1 Endometrial | 17 Esophageal Sq | 3 Ewing Sarcoma | 1
Gastric | 9 Glioma | 8 Hodgkins | 5 Kidney | 7 Liver | 20 Lung | 14 Melanoma | 28 Multiple | 6 Multiple Myeloma | 11 Nasopharyngeal
Neuroblastoma | 3 Non-Hodgkin | 2 Osteosarcoma | 10 Ovary | 14 Pancreas | 9 Pediatric ALL | 99 Prostate | 21 Testicular | 5 Thyroid | 3

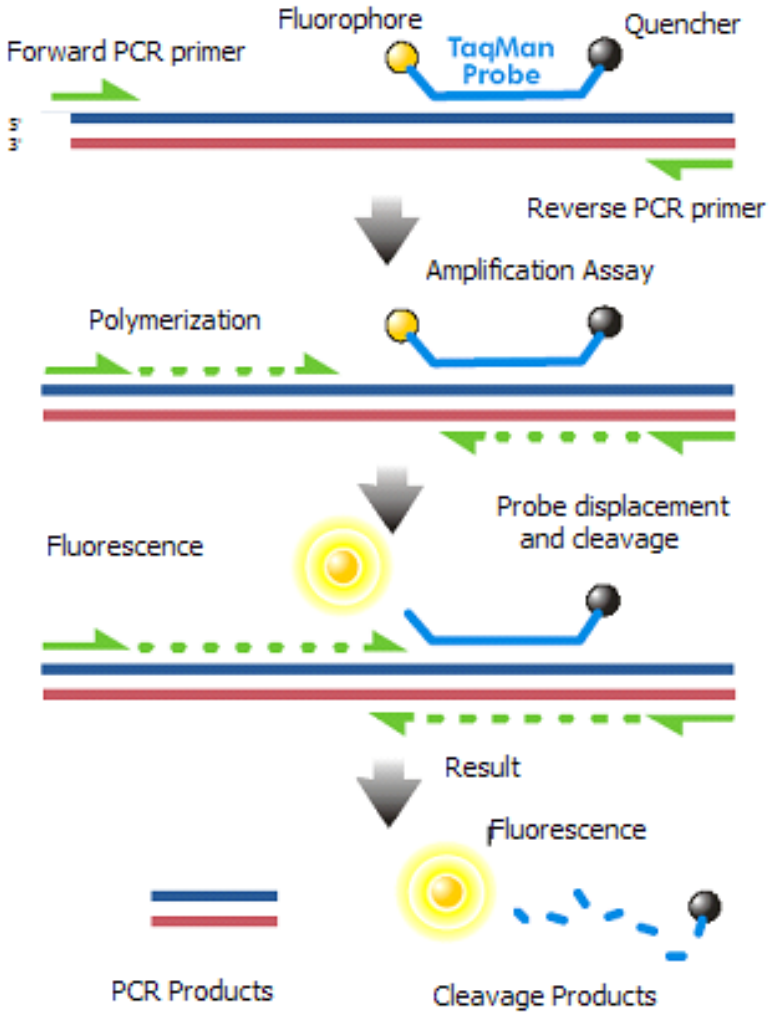
Multiple loci discovered on chr5p15.33: Telomerase Gene



Allelic Pleiotropy
Protective vs Susceptible
Gene-gene or Gene-Environment

TaqMan SNP genotyping

- Utilized for validation and extension of interesting GWAS variants
- Reasonable approach for small numbers of SNPs (<50) and large numbers of samples (assay development = $$)
- Standard locus specific forward and reverse PCR primers
- Two ~16 base probes with single base mismatches to detect SNP alleles



GWAS of Cancer Susceptibility – only the beginning

~450 distinct alleles across genome for more than 24 cancers/related traits

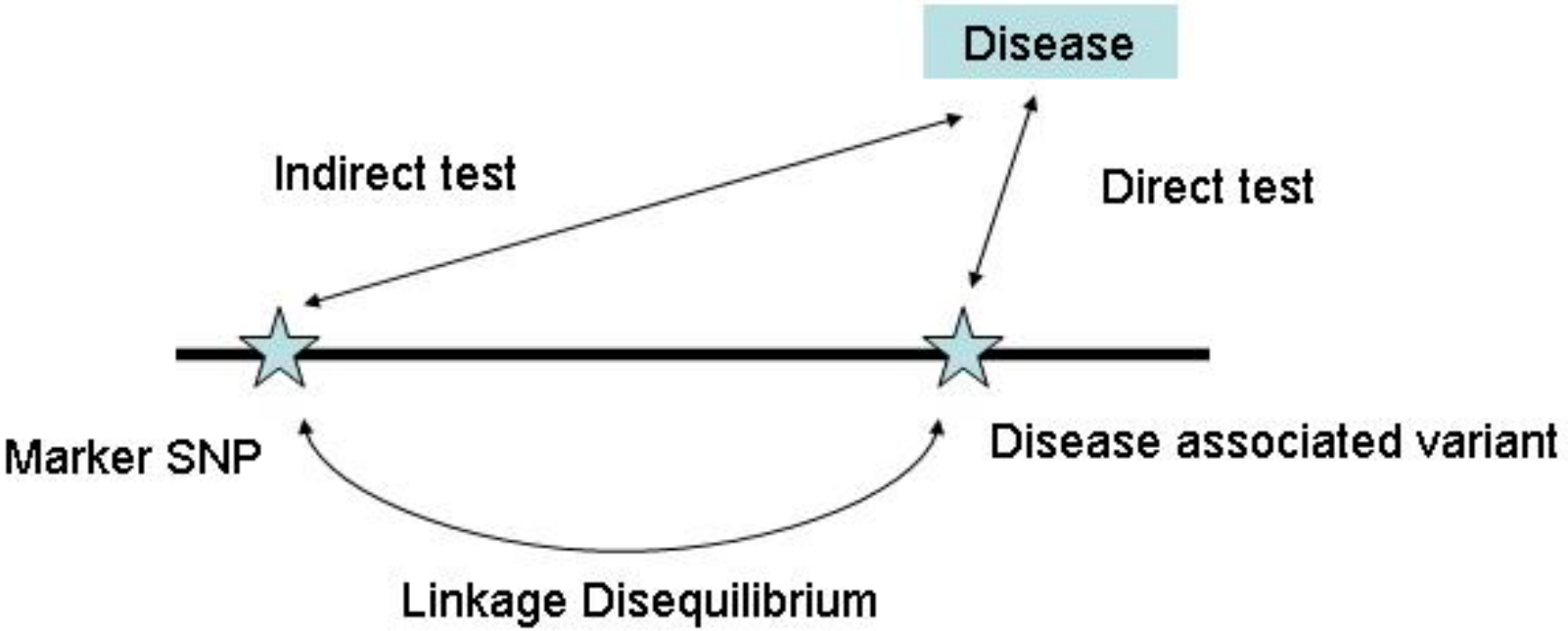
~25 'biologically explained'
<10% map to COSMIC mutations

Intervals including GWAS SNPs containing >800 genes and 25% loci are 'intergenic'

Complex genetic model for susceptibility in population

Correlation does not imply causation

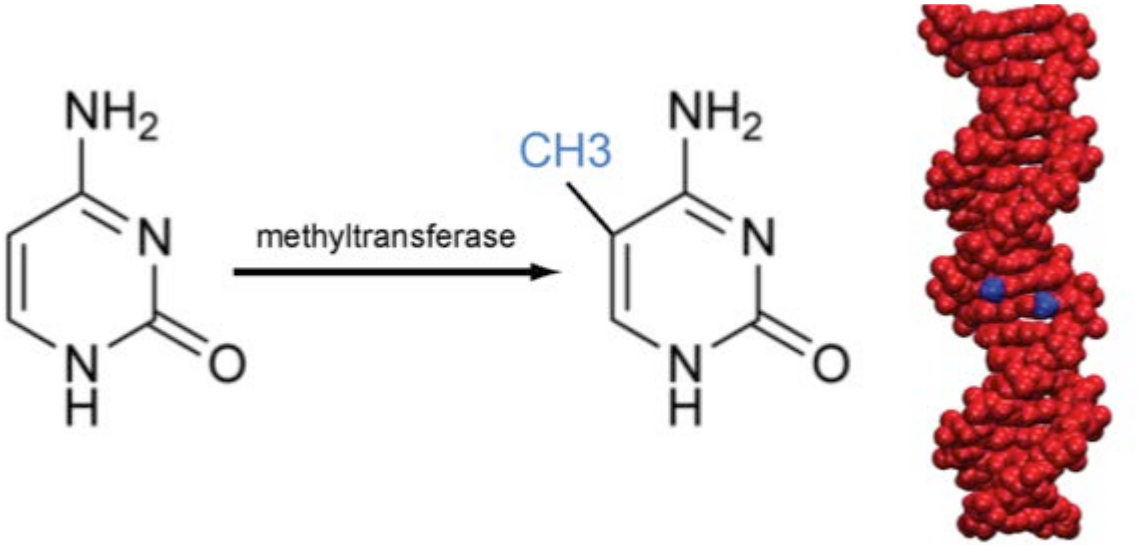
- Conventional GWAS designs provide no mechanism to distinguish statistical association from causation



Epigenetics - Methylation

Epigenetics - Methylation

- Methylation of cytosine – a covalent modification of DNA wherein the H5 hydrogen of the base is replaced by a methyl group

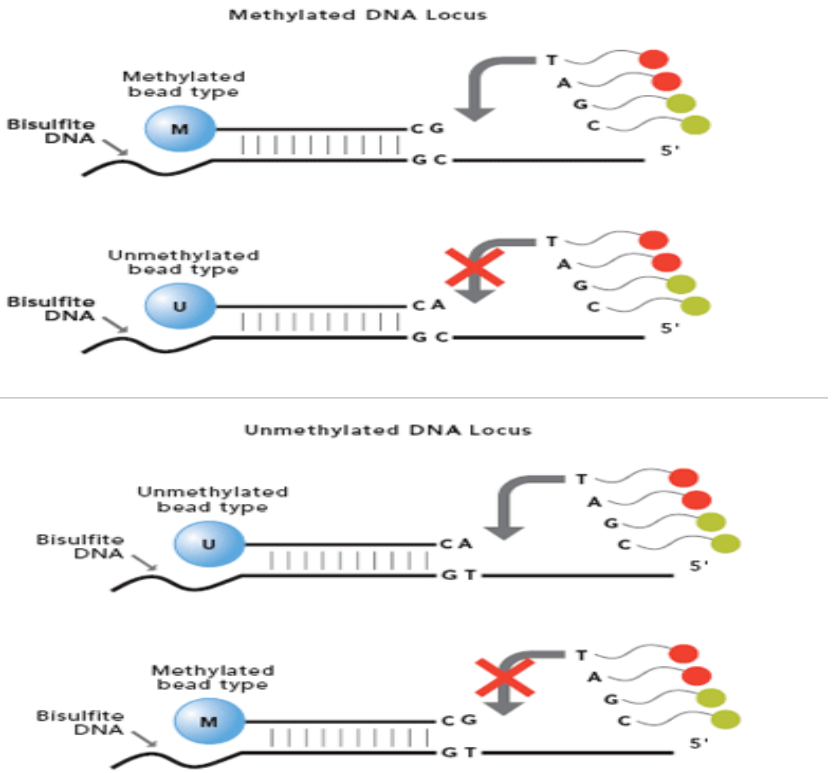


Epigenetics – Methylation-2

- DNA methylation plays a role in regulating gene expression
- Important during the developmental process
- Also important for genome stability, chromatin compaction, genome defense, and other things
- Aberrant methylation (hypermethylation or hypomethylation) may cause genes to be over- or under-expressed and has been implicated in many diseases, including cancer
  - Up-regulation of oncogenes
  - Down-regulation of tumor suppressors

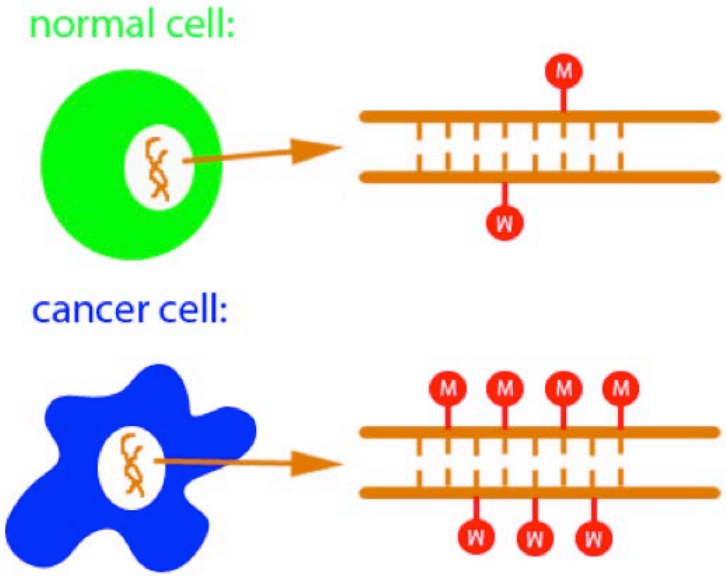
Illumina HumanMethylation450 BeadChip

- Enables epigenome-wide association studies
- Detects cytosine methylation at CpG islands based on highly multiplexed genotyping of bisulfite-converted gDNA
  - Bisulfite conversion: converts unmethylated cytosine bases to uracil, while methylated cytosine bases remain unchanged
  - Chemically differentiated loci are interrogated using two site-specific probes, one designed for the methylated locus and another for the unmethylated locus
- 99% RefSeq genes covered
  - Avg 17 CpG sites per gene region (promoter, 5'UTR, first exon, gene body, 3'UTR)



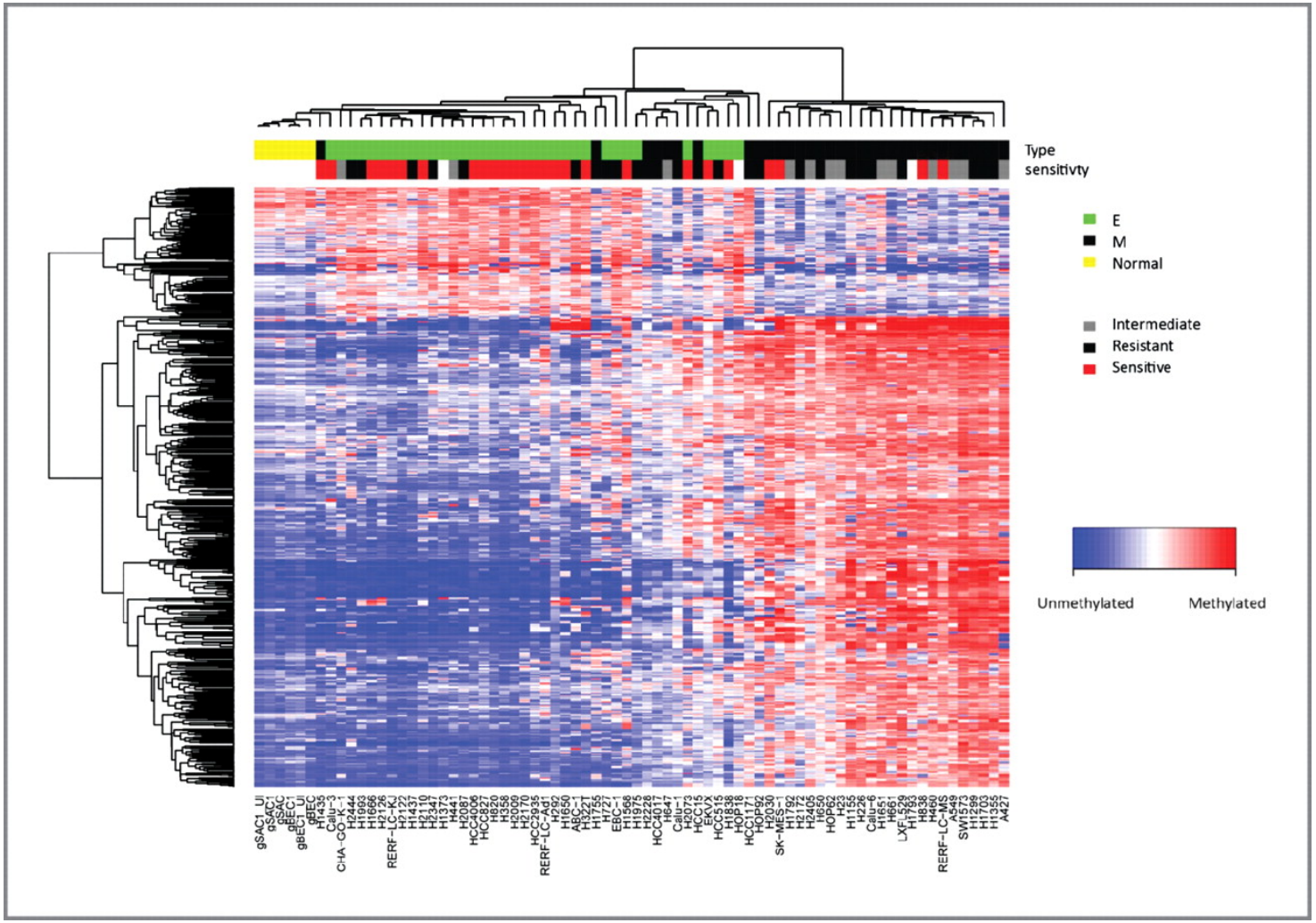
Epigenetics – Methylation use case

- Methylation in tumor vs. normal tissue from the same organ/site



- Methylation patterns in a patient's DNA before and after treatment (e.g. chemotherapy)

DNA methylation profiling delineates epithelial-like (E) and mesenchymal-like (M) NSCLC cell lines.



Kim Walter et al. Clin Cancer Res 2012;18:2360-2373

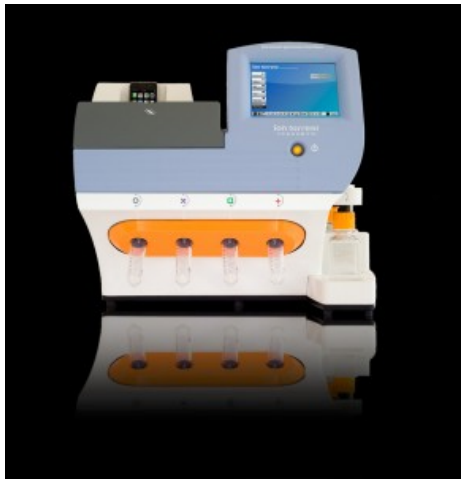
Clinical Cancer Research
AACR American Association for Cancer Research

Sequencing

Illumina HiSeq

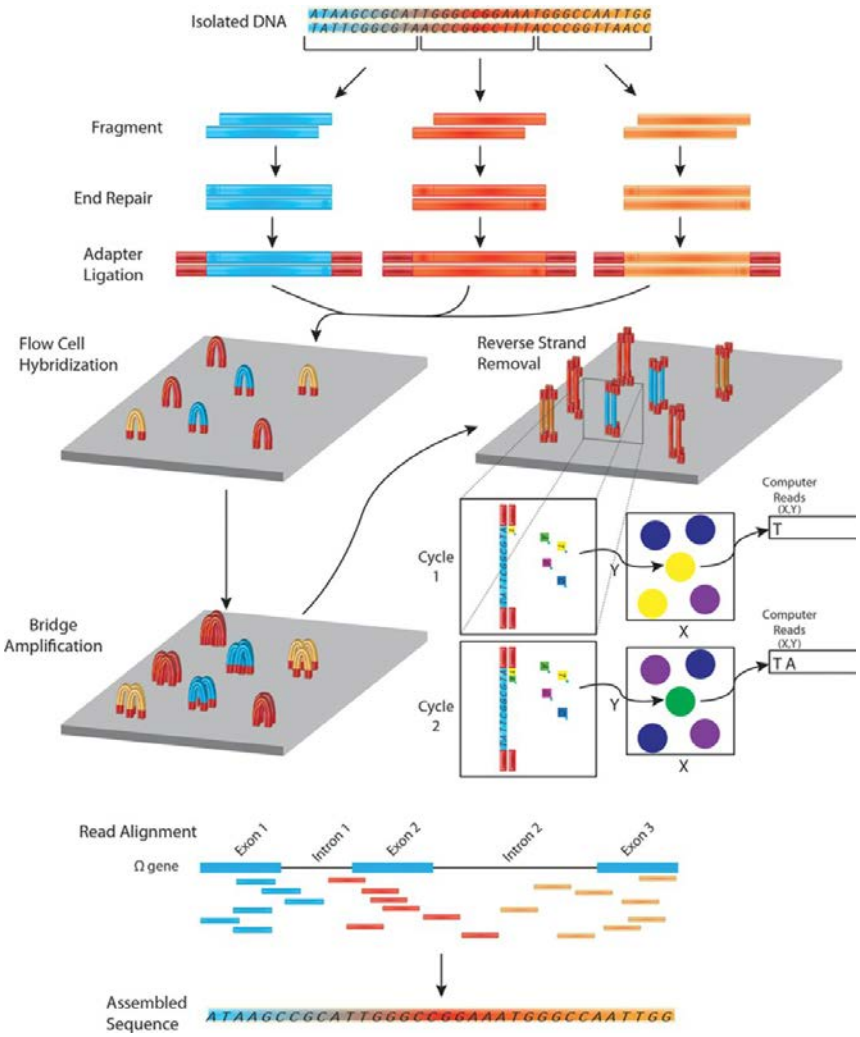
Ion Proton

Illumina MiSeq



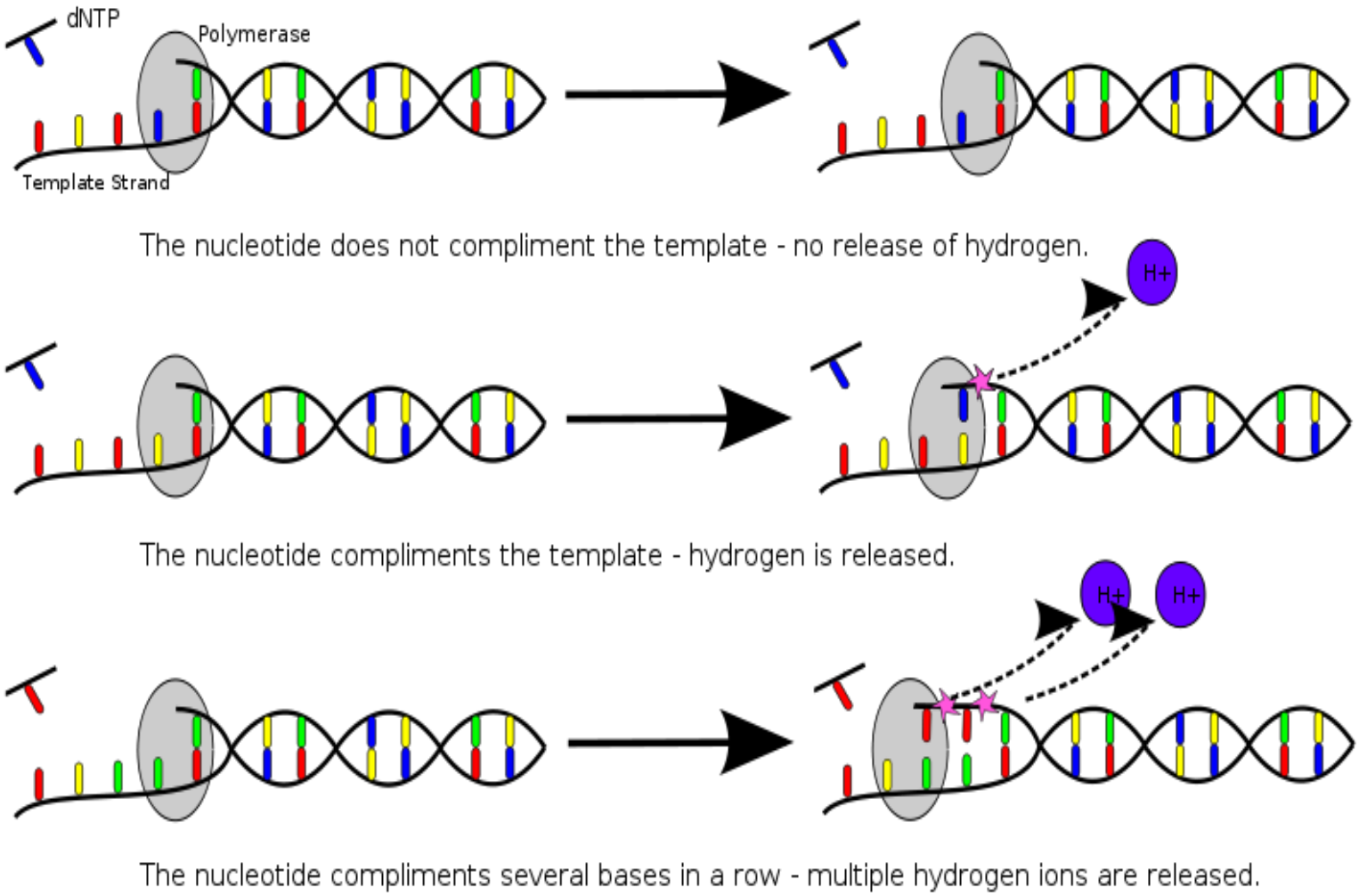
Ion Personal Genome Machine (PGM)

Overview of DNA sequencing using the Illumina platform.



Jared M. Churko et al. Circulation Research. 2013;112:1613-1623

Ion Torrent – the world's smallest pH meter



Sequencing applications

- Complex, common diseases
  - Regional sequencing for fine-mapping
    - Common variant
    - GWAS follow-up
- Rare, Mendelian diseases
  - Exome sequencing
    - Rare protein-coding variant
    - Highly penetrant
    - Families in which many members are affected
- Whole-genome sequencing
    - Function and annotation agnostic

Sequencing – use cases

- Whole-exome sequencing to look for protein-coding variants
  - Example: families
- Target-gene(s) sequencing to study only a particular gene or set of genes
  - Example: *BRCA1*/*BRCA2*
- Regional sequencing of a contiguous part of a chromosome
  - Example: GWAS follow-up
- Whole genome sequencing
  - Look everywhere in the genome; structural variation detection

Exome = Protein Coding Genome

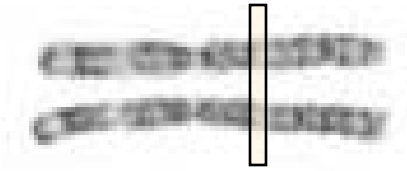
Exome = 180K exons; ~30 Mb (~1% of genome)

Advantage	Much less sequencing
Disadvantage	Miss non-coding variants

Why coding?

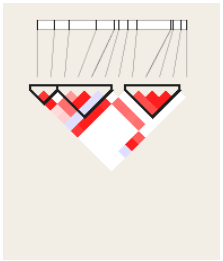
- Larger effect sizes
- More interpretable
- Easier to follow up

i. Initial association finding



← GWAS, linkage, candidate gene

ii. Exploration of known variants



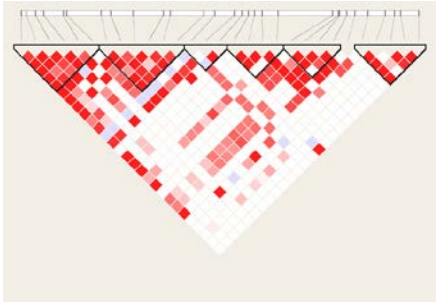
← HapMap, 1000 genomes

iii. Polymorphism detection

```
ATGGAAAGGGAGCATATTACGCACTATGAAGCGGGGGAACGA
          AGCATATTATGCACTATGAAGC
     AAGGGAGCATATTACGCACTATGAAGCGGAGGAACGA
ATGGAAAGGGAGCATATTATGCAC          GAGGAACGA
ATGGAAAGGGAGCATA          ATGAAGCGGGGGAACGA
   AAAGGGAG               TATGAAGCGGAGGAACGA
          GCATATTACGCACTATGAAG          AACGA
ATGGAAAGGGAGCAT                       GGAACGA
               TATTATGCACTATGAAGCGGAGGAACGA
ATGGA               TGCACTATGAAGCGGGGGAACGA
ATGGAAAGGGAGCATATTACGCACTAT           AACGA
```

← Characterization of all common genetic variants

iv. Comprehensive exploration of all variants



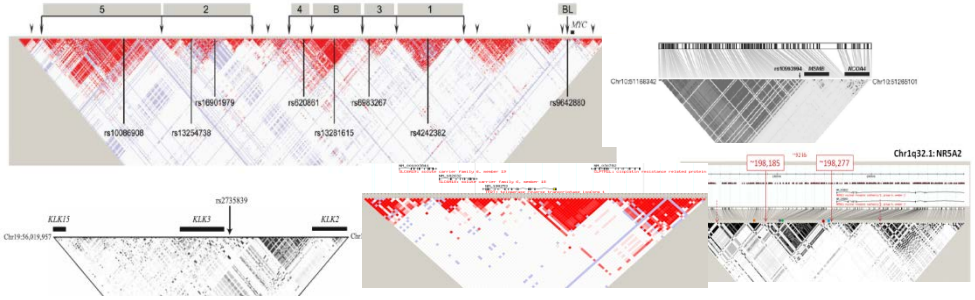
← Genotyping in large numbers of individuals

↓

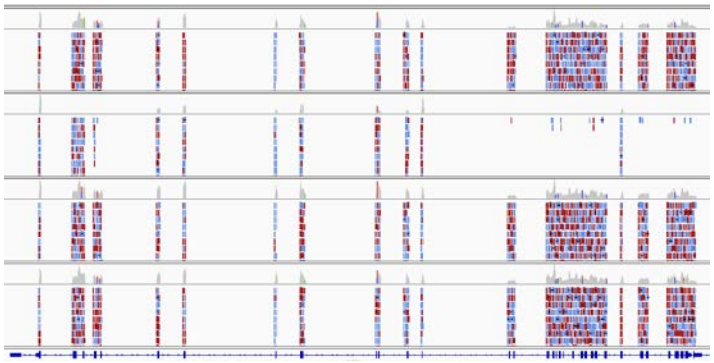
Functional characterization

Exome sequencing

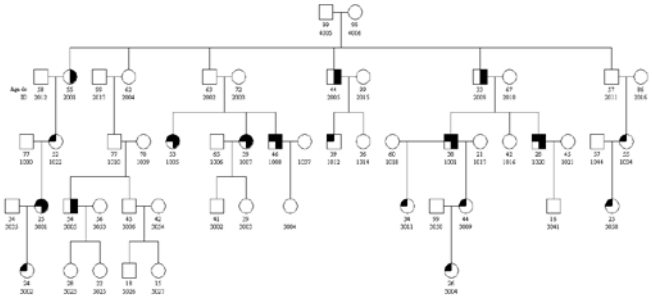
- Precursor to whole-genome sequencing
  - Cost is feasible
  - Protein-coding variants are most easily annotated and interpreted for likelihood of disease involvement
- Rare variants in rare diseases
  - Study families with extremely high disease incidence to find causal protein-altering variants
  - If not mapped by linkage methods, variants may be novel and not co-localized among families
  - Assumes virtually complete penetrance
- Design has power to find different variants than GWAS



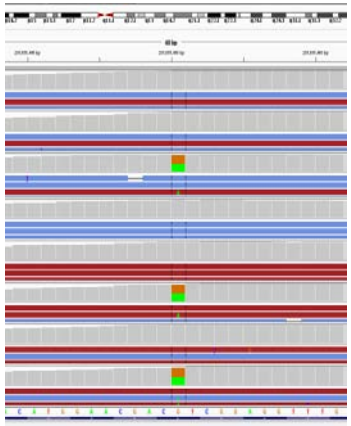
Regional GWAS and linkage follow-up



Candidate genes



Whole exome



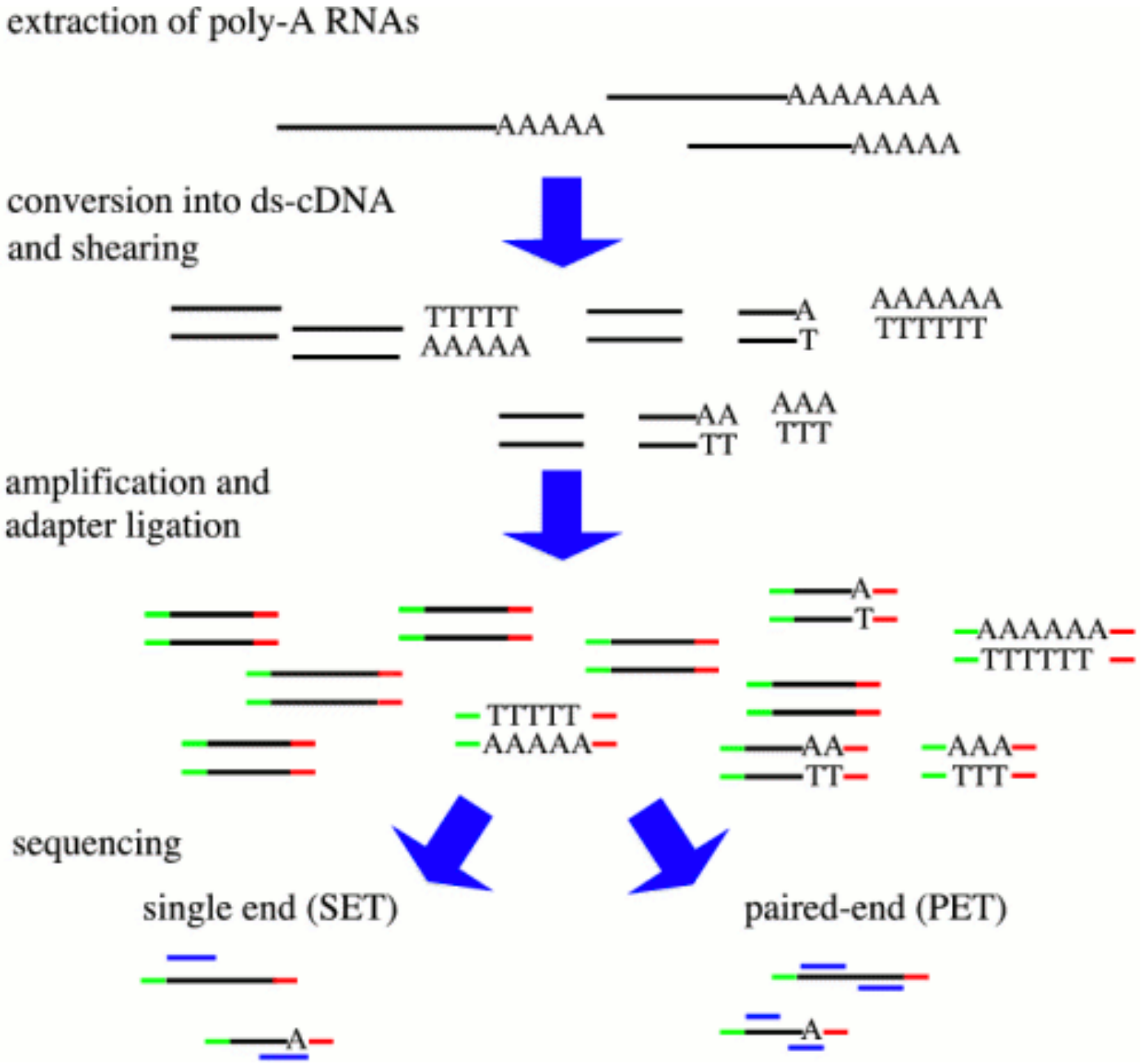
Technical Validation For Exome

Whole genome

Transcriptome - RNASeq

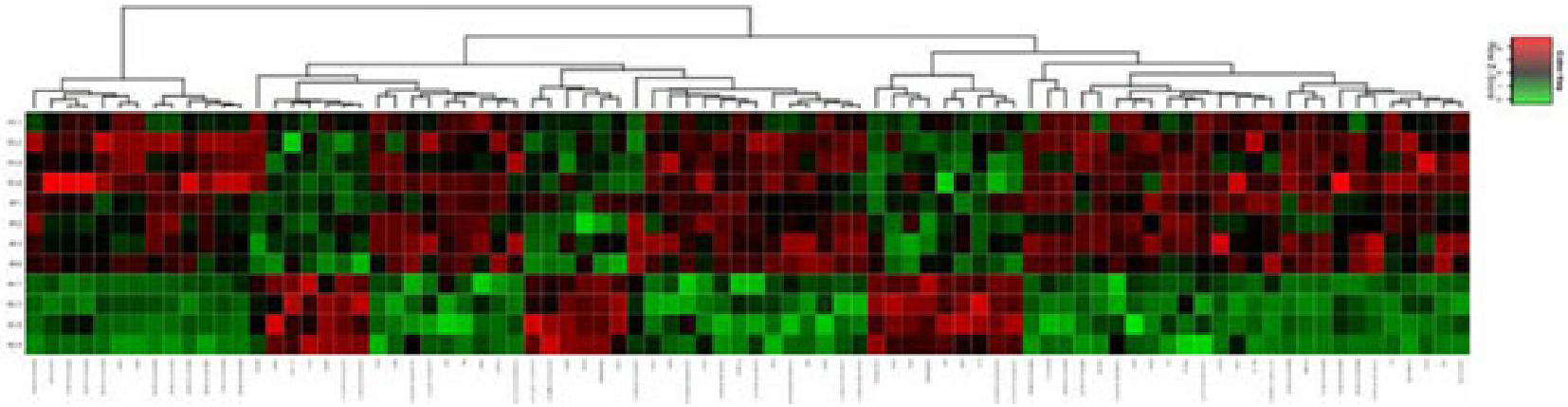
Transcriptome

- The set of all RNA transcripts in a cell or population of cells
  - mRNA
  - rRNA
  - tRNA
  - Non-coding RNAs, including miRNAs
- Tissue- and cell-type specific
- We can look at the whole transcriptome, but can also only target certain genes


extraction of poly-A RNAs
AAAAAAA
AAAAA
AAAAA
conversion into ds-cDNA
and shearing
TTTTT
AAAAA
A
T
AAAAAA
TTTTTT
AA
TT
AAA
TTT
amplification and
adapter ligation
A
T
AAAAAA
TTTTTT
TTTTT
AAAAA
AA
TT
AAA
TTT
sequencing
single end (SET)
paired-end (PET)
A
A

Transcriptome – use cases

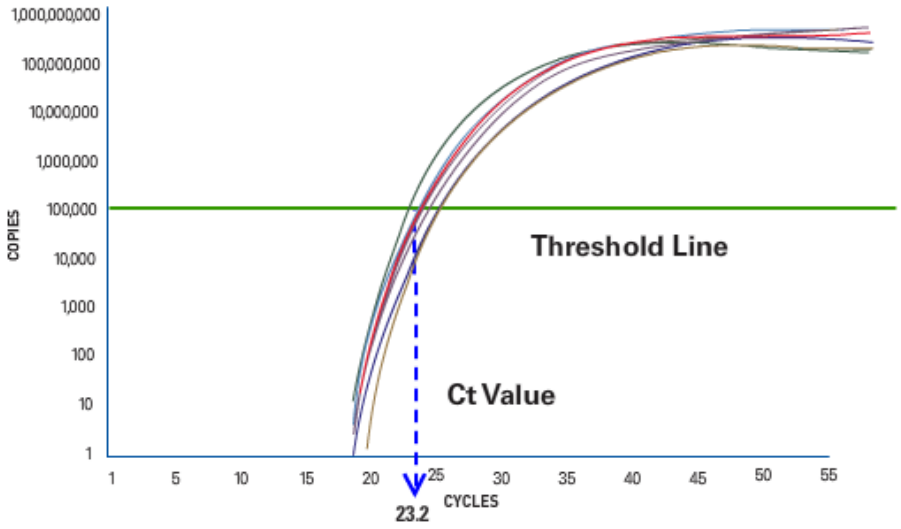
- Examine differential expression in genes from paired tumor/normal samples from the same individual
- Do different subtypes of breast cancer tumors have different expression signatures?



- Do we see novel fusion transcripts in tumors?
  - Example: *PAX8-PPARG* fusion in follicular architecture-follicular thyroid carcinoma

Quantitative (real-time) PCR assays

- Quantitation by interrogating PCR cycle where amplification passes threshold
- Can be measured via absolute (e.g. against standard curve) or relative (e.g. against endogenous control)
- Can be used for:
  - Gene expression studies
  - Copy number variation
  - Absolute quantitation
  - SNP genotyping
  - Pathogen detection



Telomeres

Telomeres

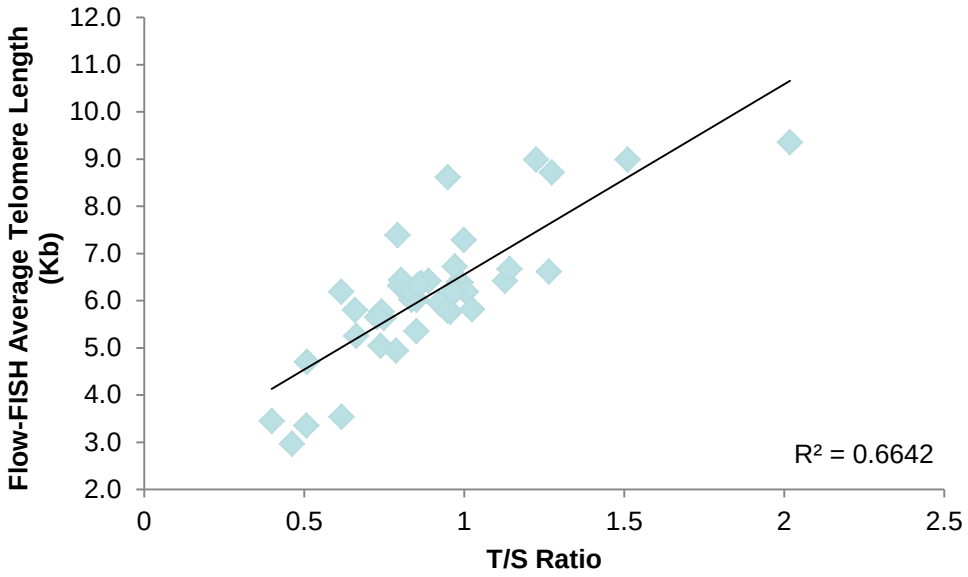
- Ends of chromosomes
- Repetitive nucleotide sequences (TTAGGG × ~2,500 in humans)
- When cells divide, enzymes that duplicate chromosomes cannot reach the end, so the chromosome is shortened – association with aging
- Telomeres serve as “buffers” to protect the remainder of the chromosome
- Telomeres can be replenished by *telomerase reverse transcriptase* (*TERT*)
- There are other genes that function, along with *TERT*, in telomere maintenance

Telomeres - 2

- Telomere dysfunction is commonly acquired during the process of tumor development
  - Short telomeres can lead to genomic instability, chromosome loss as well as to translocations
  - Shortened telomeres have been observed in many cancer types, even in the blood DNA of the patients
- Individuals with dyskeratosis congenita are born with short telomeres
  - Rare germline variants in key telomere maintenance genes

Telomere Length Assay

- Relative average telomere length determination by qPCR
- Single-plex Telomere assay (T) and Autosomal single copy gene (S)
- Concentration of each target extrapolated from a standard curve
- Resulting T/S ratios are proportional to average telomere length
- Pilot studies show T/S ratios are strongly correlated to Flow-FISH cytometry results



Telomere Length Assay – use cases

- Compare telomere lengths between people with cancer and people without
- Compare telomere length between tumor and normal tissue
- Used as a diagnostic assay for people suspected of having dyskeratosis congenita

Integrative analyses

The next goal

Now that we've generated all of this data…

- We can analyze it by specific platform (e.g. SNP data) for its intended purpose
- How can we combine and look at it all together?
  - Could SNP genotype + methylation status + RNA expression be linked and important? Definitely.
  - How could SNPs in the genome predict telomere length and therefore associate with disease?
  - How could a SNP found by exome sequencing affect RNA expression or alternative transcripts?
  - And on and on …

Future considerations for study planning

- Given all of the technologies and approaches available, how can we best combine the data to get the most meaningful results?
  - SNP associations + methylation + RNASeq = ?
- Sample collections should be thought out carefully
- We are living in exciting times, technologically-speaking

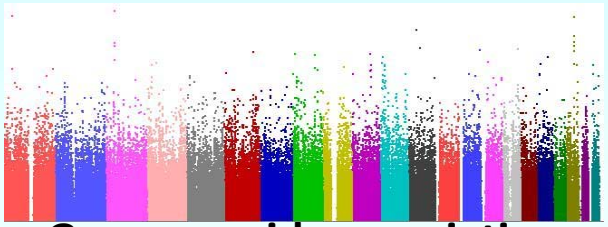
Conclusions

- There are now many ways to biologically approach our questions of disease
- Multiple technologies and assays are available in order to best explore biological questions
- Integrative analyses will be key to solving the puzzles we are facing

If there's time…

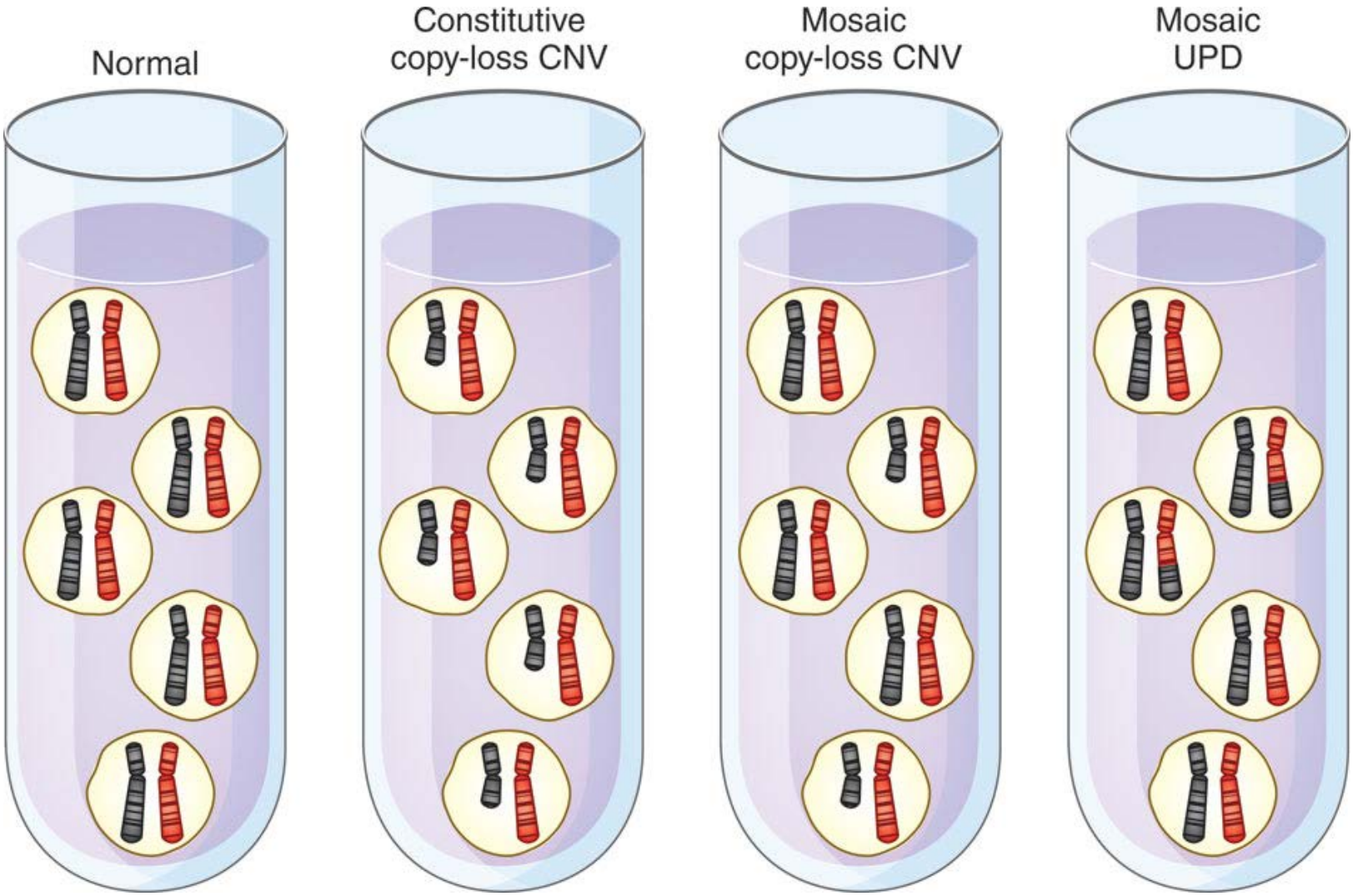
Exciting new stuff.

Unexpected findings



Genome-wide association studies

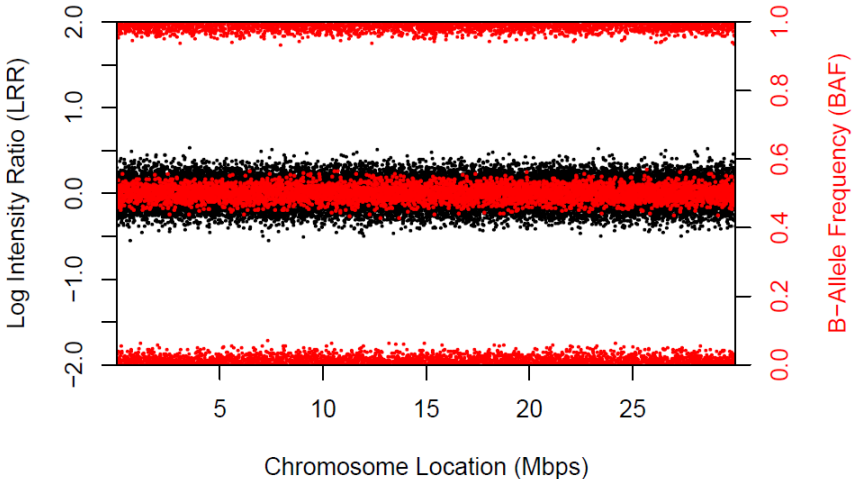
Somatic mosaicism



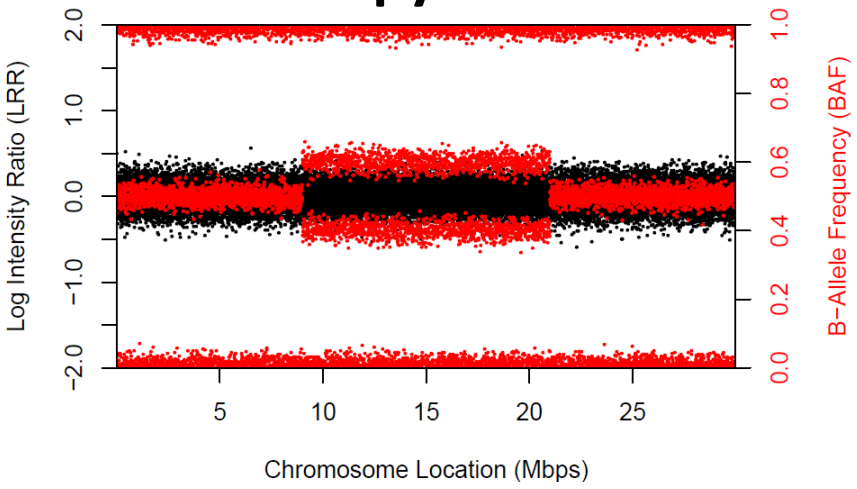
Nature Genetics 44, 614–616 (2012) doi:10.1038/ng.2311

Examples

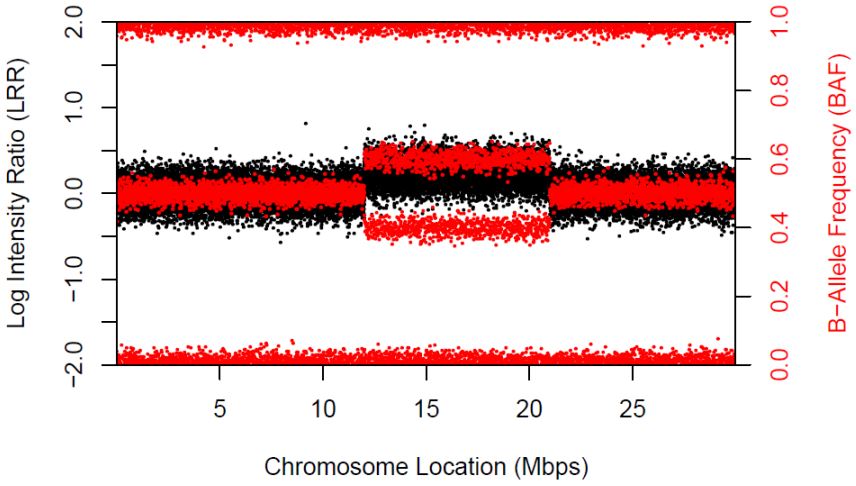
Normal



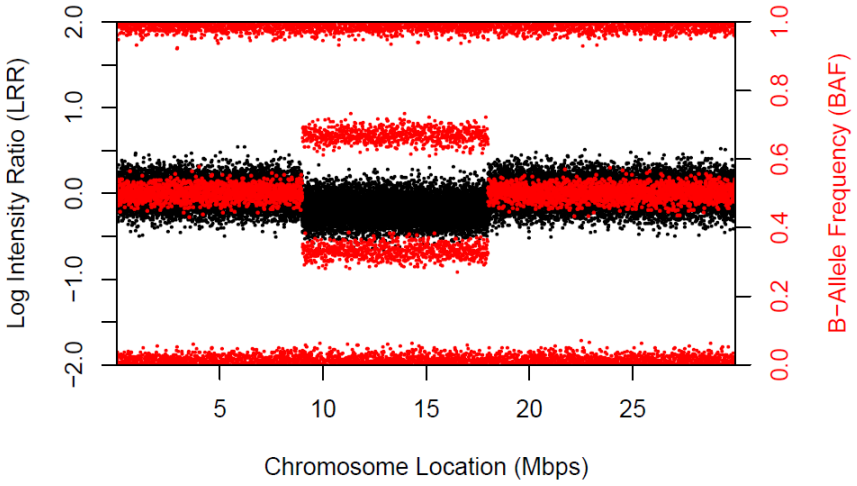
Mosaic Copy-Neutral LOH



Mosaic Gain



Mosaic Loss

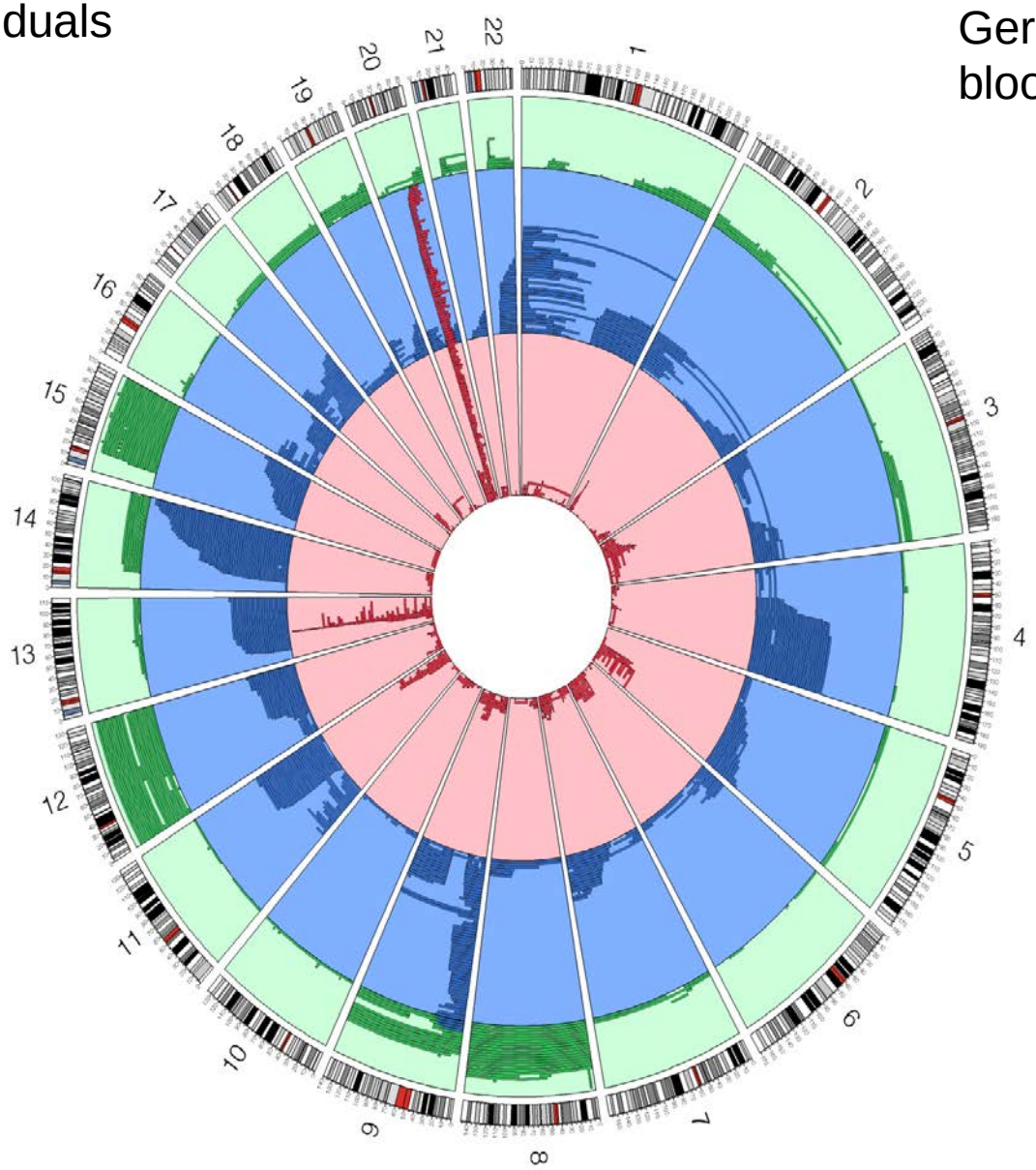


- Validated with FISH, MLPA, and microsatellite analysis

Circos plot of mosaic events

127,179 individuals

Germline DNA from blood or buccal



Green = gain
Blue = CNLOH
Red = loss

1,334 events

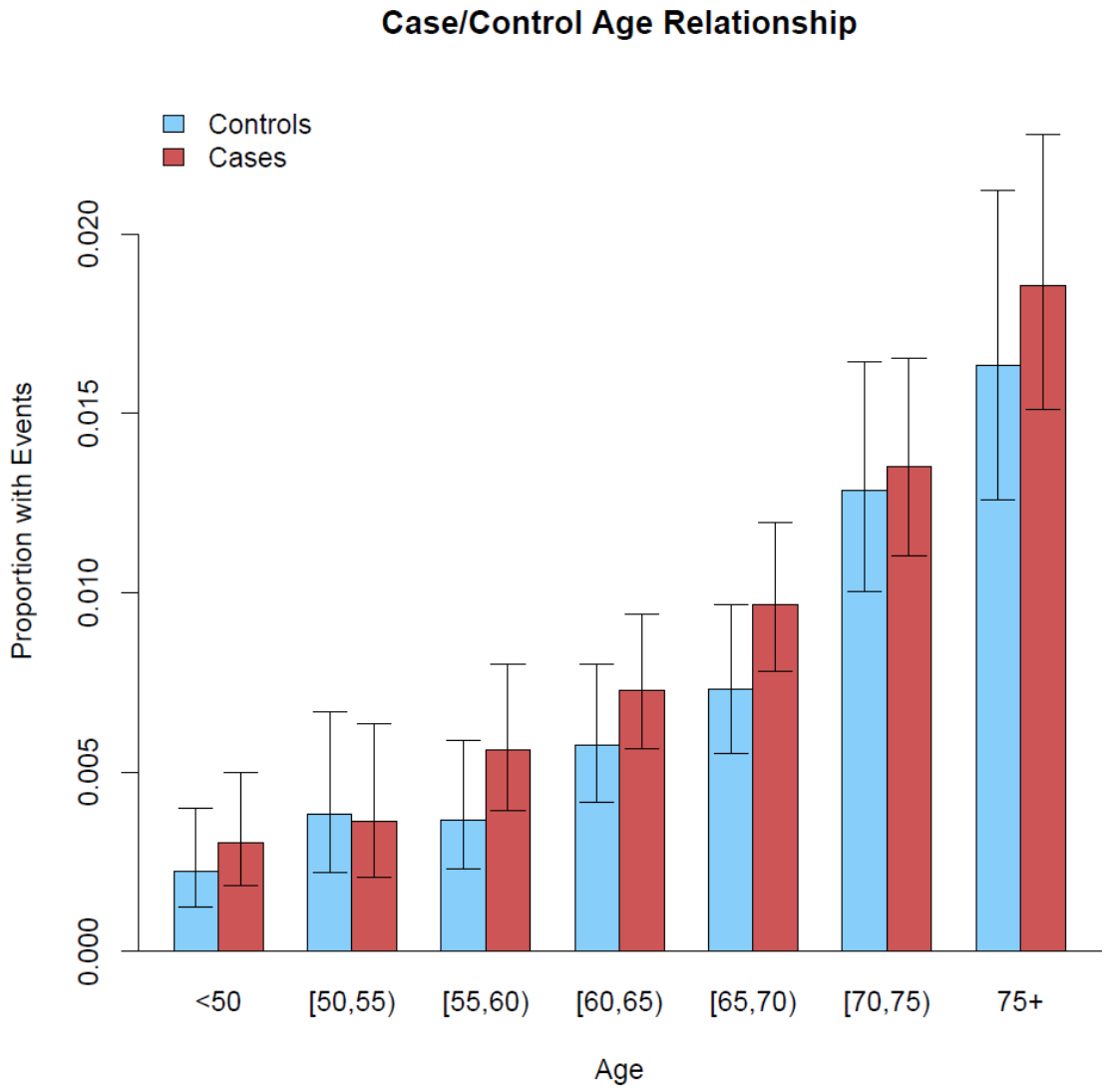
Jacobs et al 2012

Increasing Frequency with Age - autosomes

Study	P value
GENEVA	2.0×10^{-16}
TGSI	4.8×10^{-8}
TGSII	5.3×10^{-4}

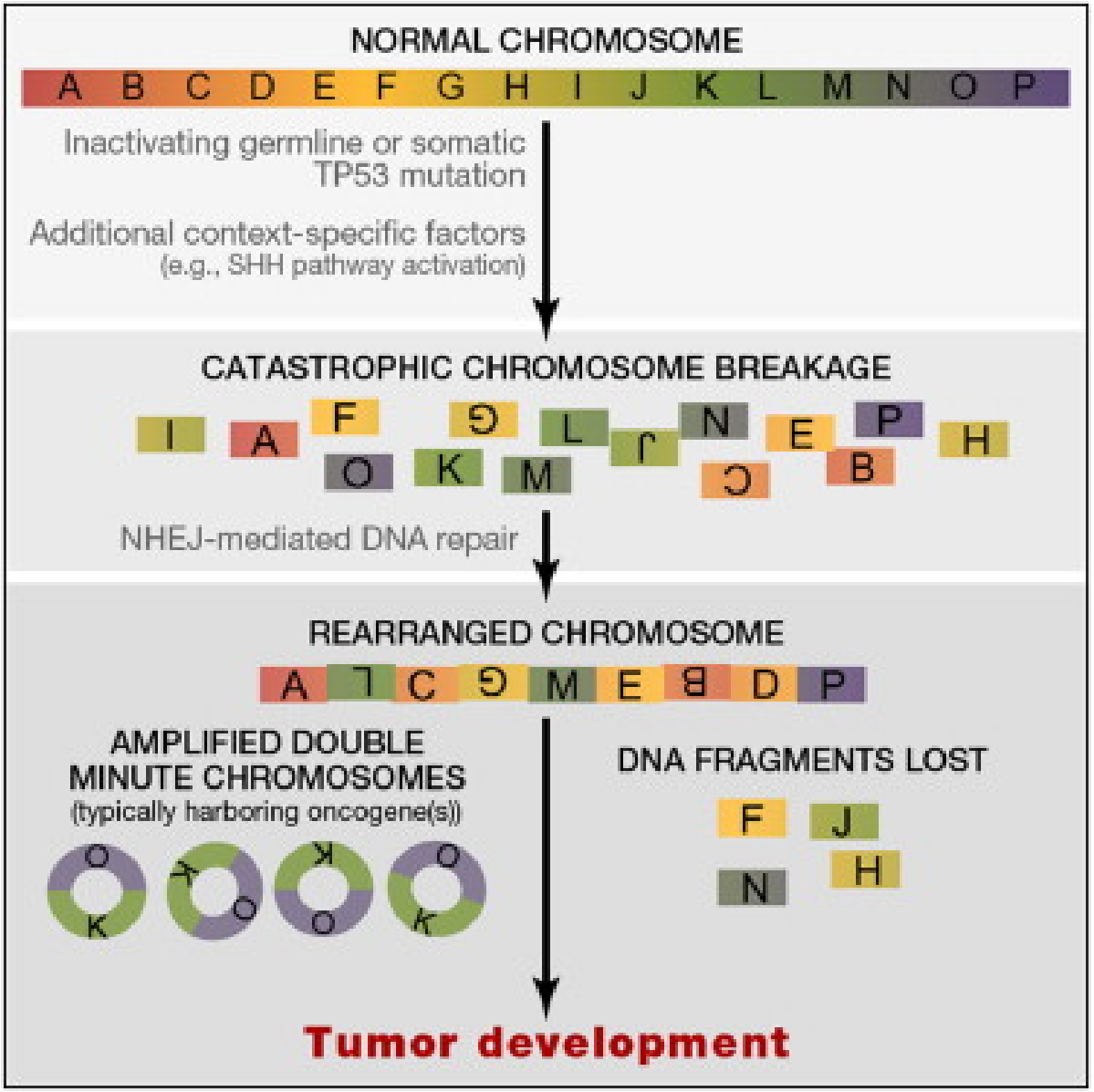
Pooled	P value
TGSI + TGSII	4.32×10^{-13}

127,179 individuals



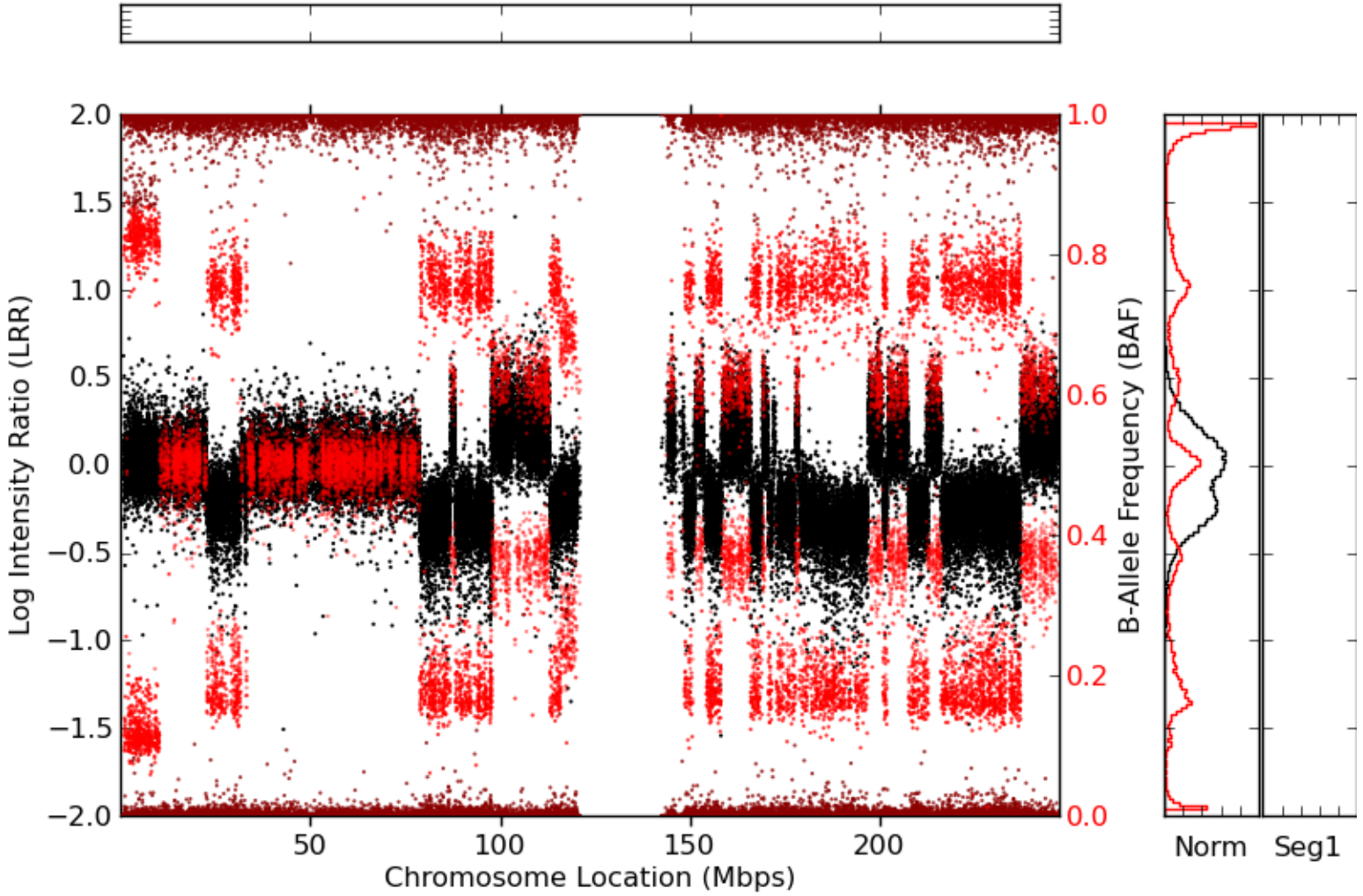
*Pooled estimate adjusted for sex, 5-year age group, cancer subtype, and study indicator; events >2Mb.

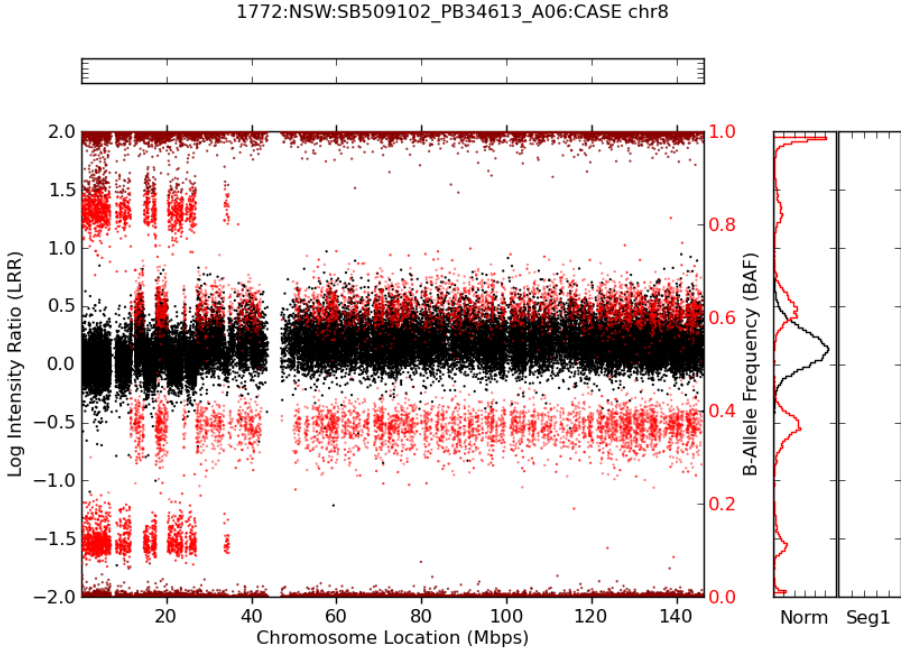
Chromosome shattering / chromothripsis / chromoplexy

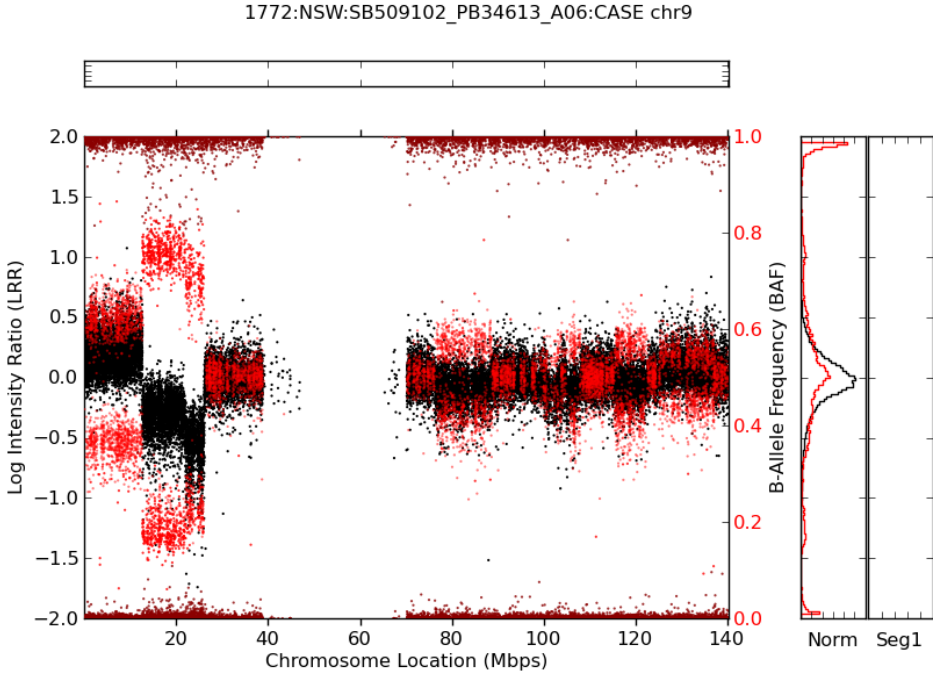


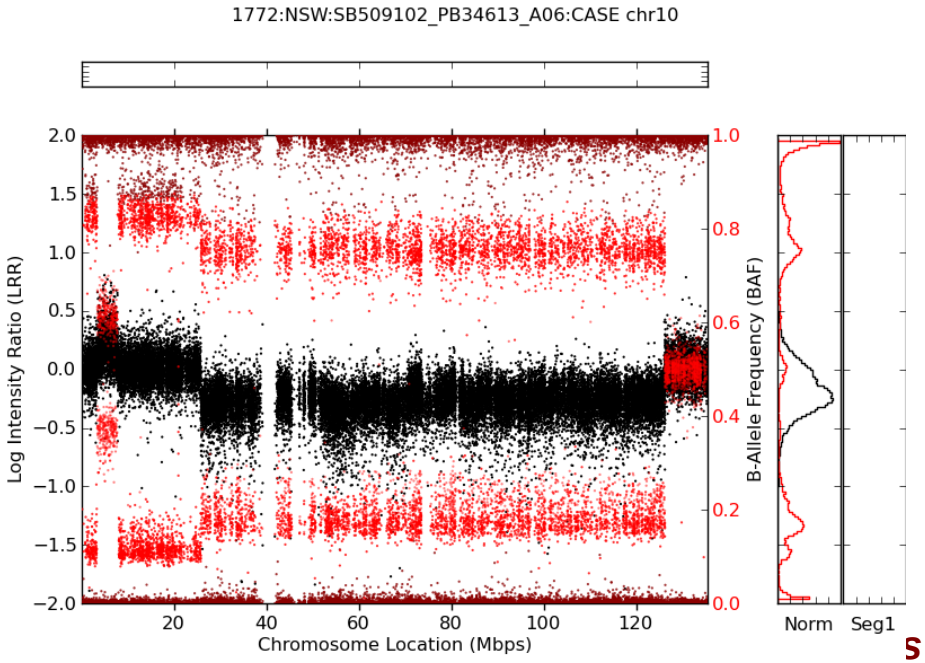
Rausch et al. Cell 2012 148(1):59-71

“Chromosome shattering”

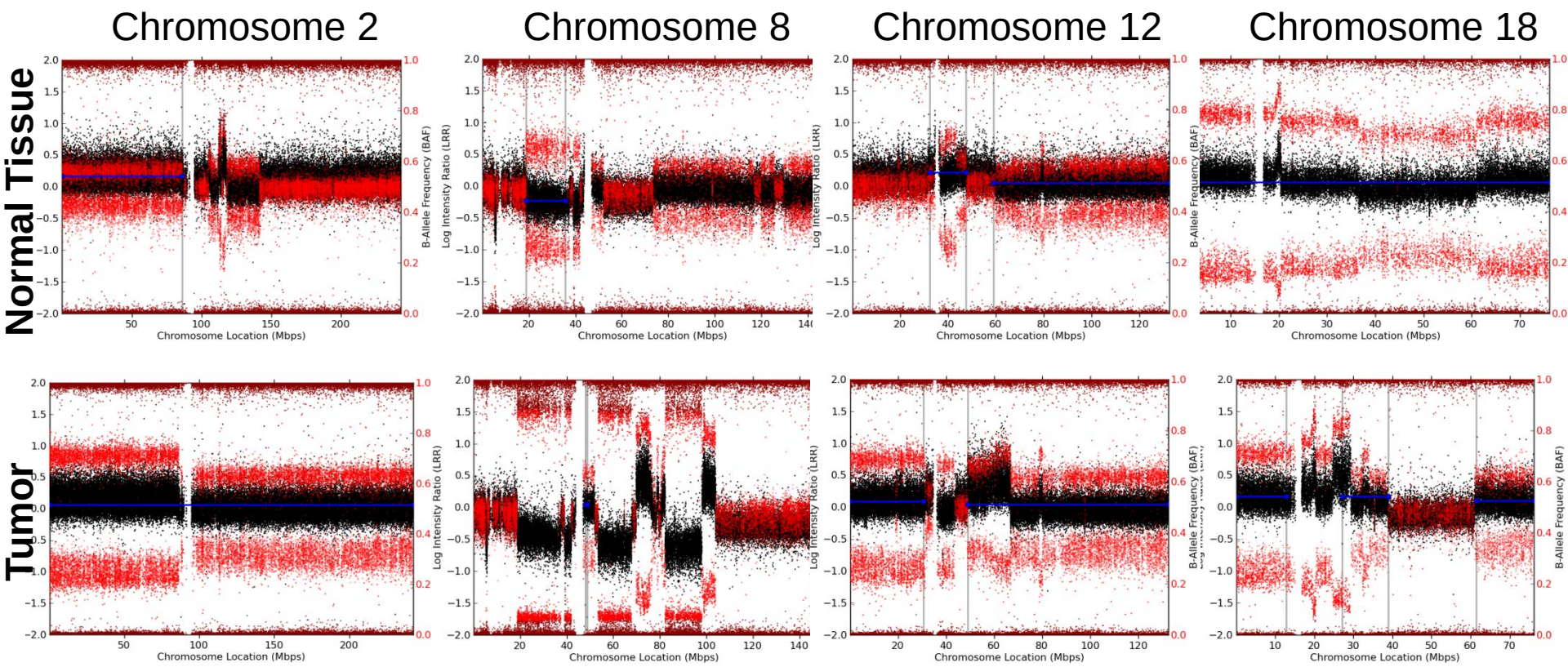



1772:NSW:SB509102_PB34613_A06:CASE chr8
Log Intensity Ratio (LRR)
B-Allele Frequency (BAF)
Chromosome Location (Mbps)
Norm
Seg1


1772:NSW:SB509102_PB34613_A06:CASE chr9
Log Intensity Ratio (LRR)
B-Allele Frequency (BAF)
Chromosome Location (Mbps)
Norm
Seg1


1772:NSW:SB509102_PB34613_A06:CASE chr10
Log Intensity Ratio (LRR)
B-Allele Frequency (BAF)
Chromosome Location (Mbps)
Norm
Seg1

TCGA-4-1371 (Ovarian)



- Events present in normal tissue that are not in tumor tissue (chr 2)
- Events arise in tumor tissue that are not present in normal tissue (chr 8,12,18)

Questions and Answers

U.S. Department of Health and Human Services
National Institutes of Health | National Cancer Institute
www.dceg.cancer.gov/RadEpiCourse
1-800-4-CANCER
Produced May 2015