Genomics Technologies Suitable for Population Studies

Michael Snyder

December 18, 2007

GGTTCCAAAAGTTTATTGGATGCCGTTTCAGTACATTTATCGTTTGCTTTGGATGCCCTAATTAAAAGTGA
CCCTTTCAAACTGAAATTCATGATACACCAATGGATATCCTTAGTCGATAAAATTTGCGAGTACTTTCAAA
GCCAAATGAAATTATCTATGGTAGACAAAACATTGACCAATTTCATATCGATCCTCCTGAATTTATTGGCG
TTAGACACAGTTGGTATATTTCAAGTGACAAGGACAATTACTTGGACCGTAATAGATTTTTTGAGGCTCAG
CAAAAAAGAAAATGGAAATACGAGATTAATAATGTCATTAATAAATCAATTAATTTTGAAGTGCCATTGTT
TTAGTGTTATTGATACGCTAATGCTTATAAAAGAAGCATGGAGTTACAACCTGACAATTGGCTGTACTTCC
AATGAGCTAGTACAAGACCAATTATCACTGTTTGATGTTATGTCAAGTGAACTAATGAACCATAAACTTGG
TCATTCG

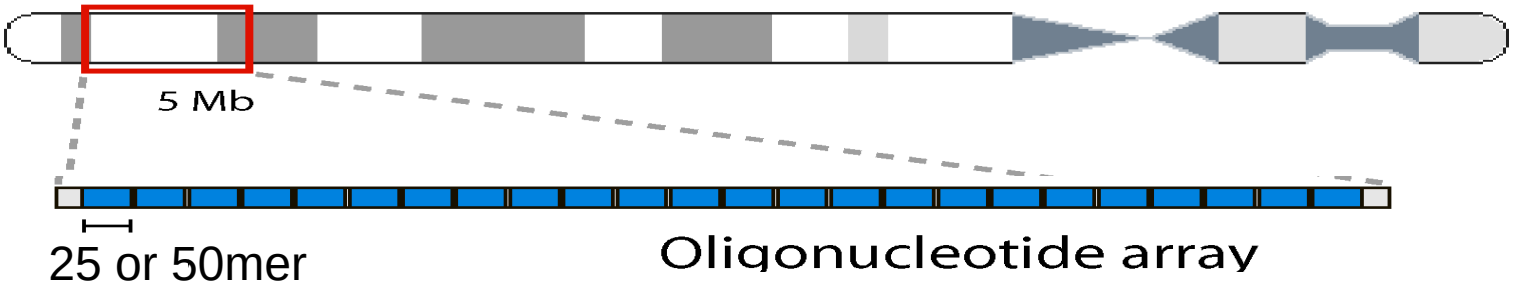
Genomics Technologies

1) Mapping DNA sequence variation

2) Mapping transcripts

3) Mapping regulatory information

Genome Tiling Arrays



Massively Parallel Sequencing

AGTTCACCTAAGA…

CTTGAATGCCGAT…

GTCATTCCGCAAT…

Human Variation: SNP Mapping

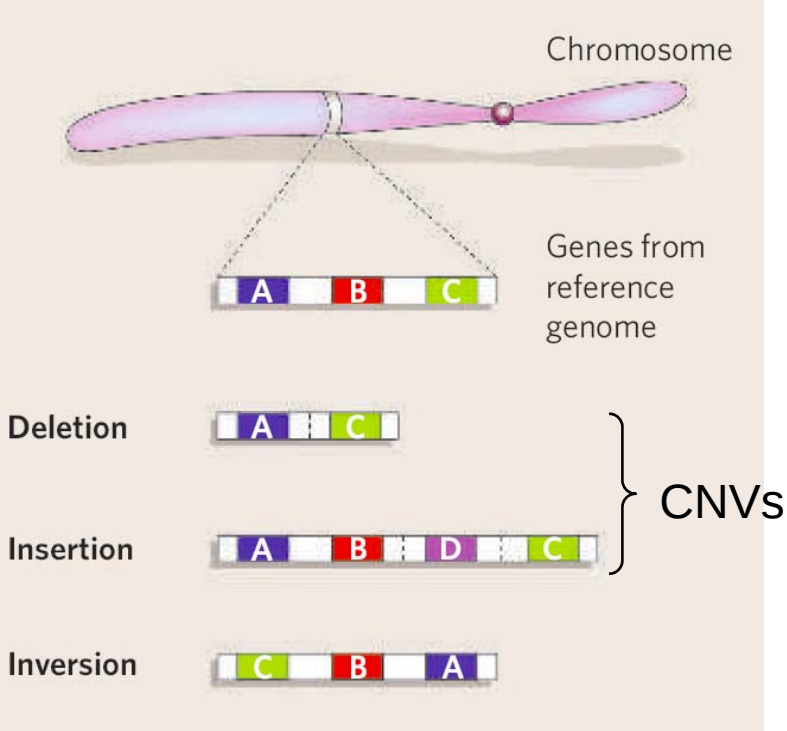
Methods:

Affymetrix 6.0
900K SNPs + CNVs

Illumina
1M SNPs + CNVs

Sequencing based
Methods

Mapping Structural Variation in Humans



- Thought to be Common
  12% of the genome
  (Redon et al. 2006)

- Likely involved in phenotype variation and disease

- Most methods for detection are low resolution (>50 kb)




High Resolution Comparative Genomic Hybridization

Array Based Methods:

BAC arrays

Affy & Illumina ~50kb

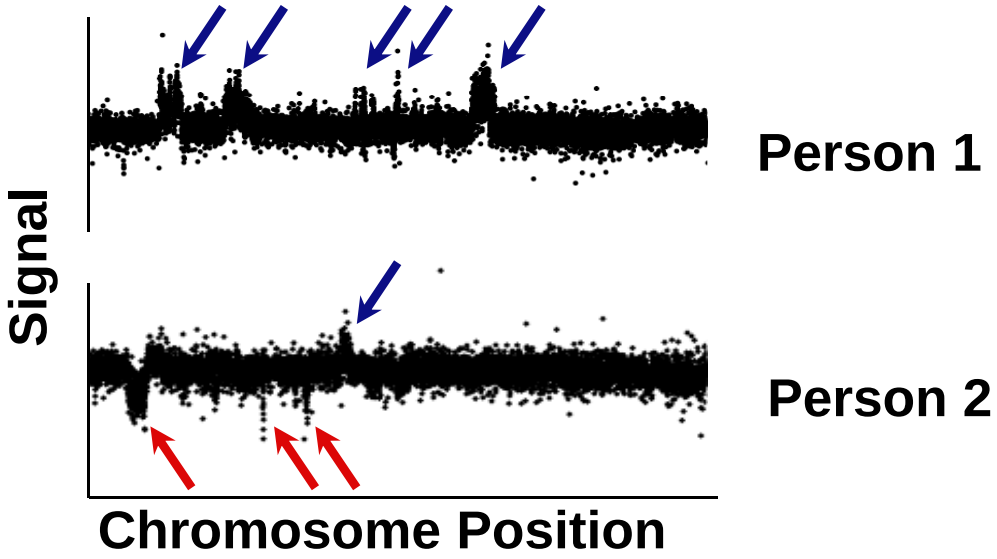
Nimblegen 400 K Array

1 array (>60kb)

8 array set (>10 kb?)

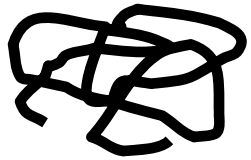
2.1 M feature arrays

1 array (>5 kb?)



Urban et al. (2006) PNAS

High Resolution-Paired-End Mapping (HR-PEM)



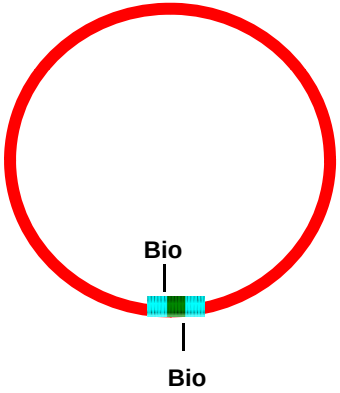
Genomic DNA

Shear to 3 kb

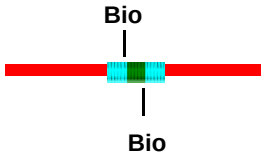
Adaptor ligation

Fragments

Circularize



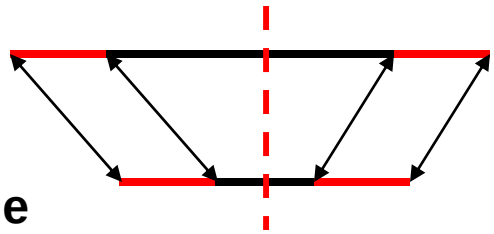
Random Cleavage



200-300bp

454 **Massively Parallel Sequencing**

(250bp reads, 400 000 reads/run)



Map paired ends to human reference genome

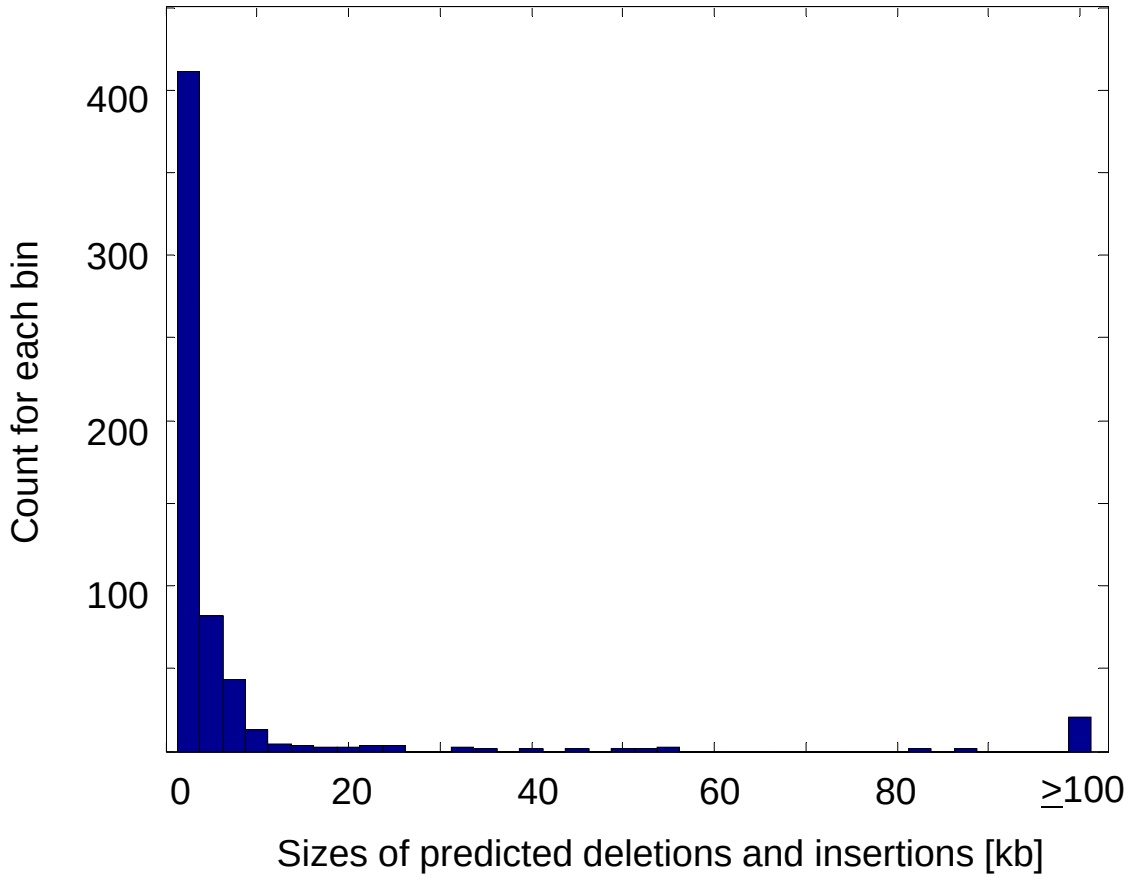
Summary of PEM Results

	NA15510 (European?, female)	**NA18505 (Yoruba, female)**
# of sequence reads	> 10 M.	> 21 M.
Paired ends uniquely mapped	> 4.2 M.	> 8.6 M.
Fold coverage	~ 2.1x	~ 4.3x
Predicted Structural Variants*	**478**	**839**
Indels	*427*	*758*
Inversion breakpoints	*51*	*81*
Estimated total variants* genome-wide	**759**	**902**

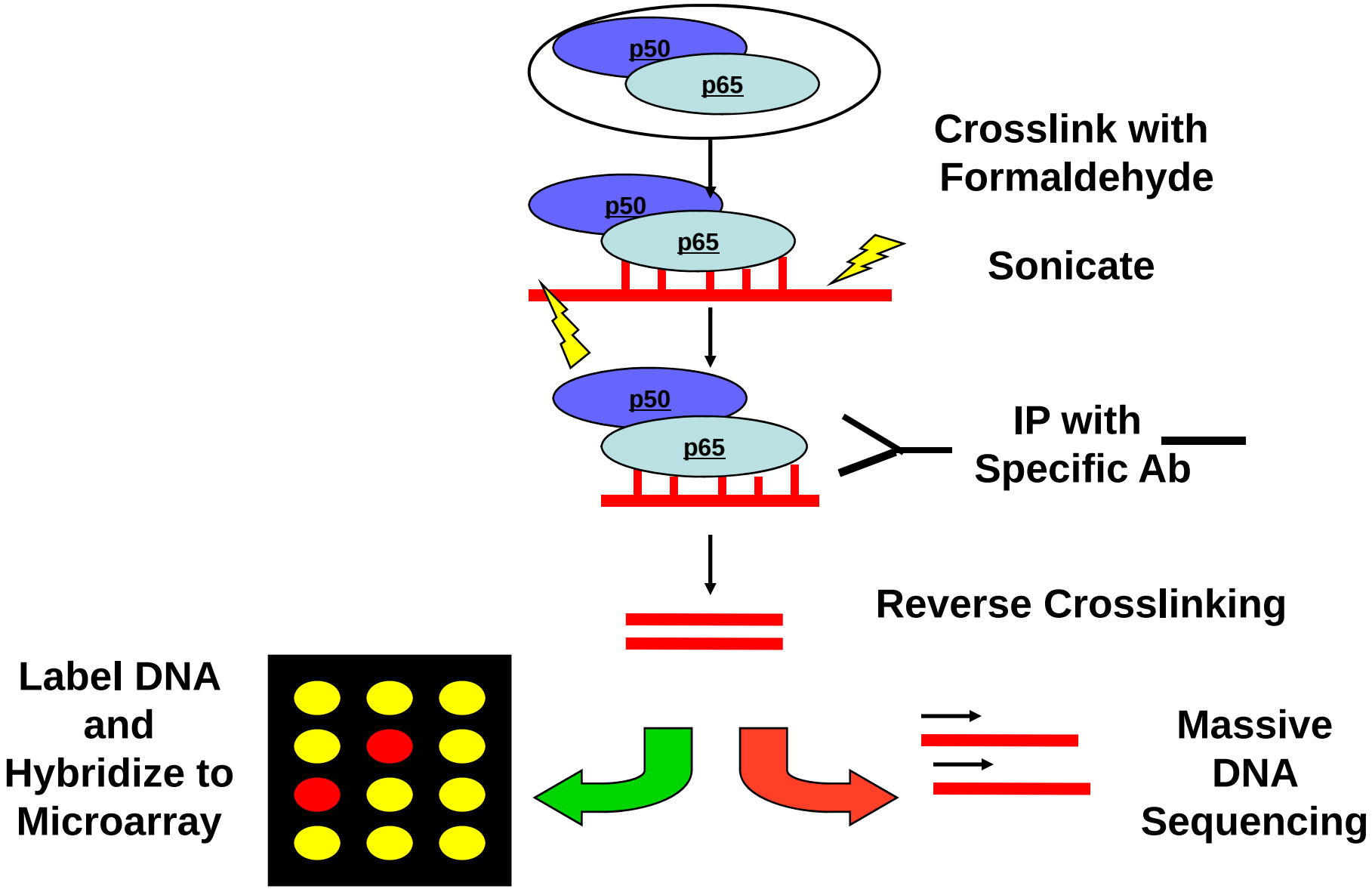
*at this resolution

Size distribution of Structural Variants

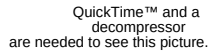
[NA15510, excluding complex rearrangements, and inversion breakpoints]

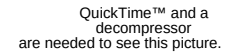


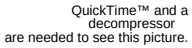
Chromatin Immunoprecipitation

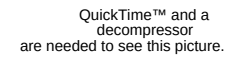


Pol II: Chromosome 22
Summary of Tracks









HeLa S3 Rep1

HeLa S3 Rep2

HeLa S3 Rep3

HeLa S3 Input

HEK293 Rep1

HEK293 Rep2

HEK293 Rep3

HEK293 Input

K562 Rep2

K562 Rep3

K562 Input

NB4 Rep1

NB4 Rep2

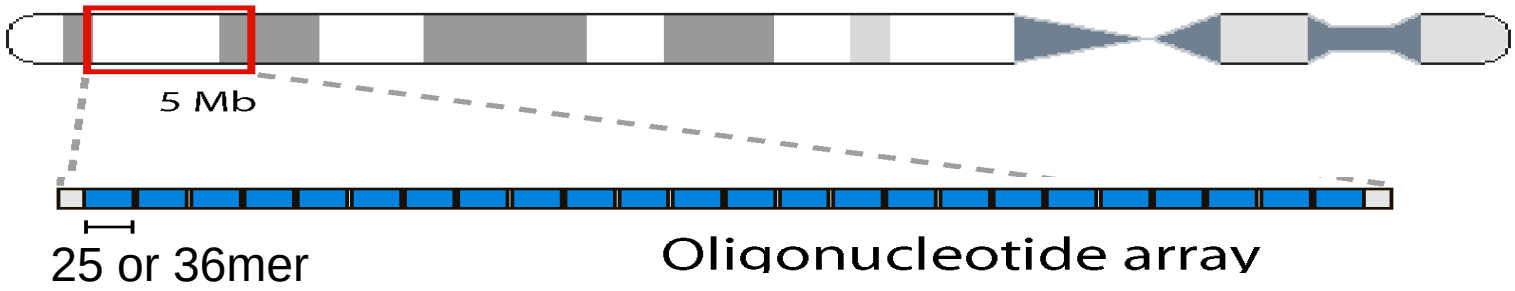
NB4 Rep3

Epigenomics

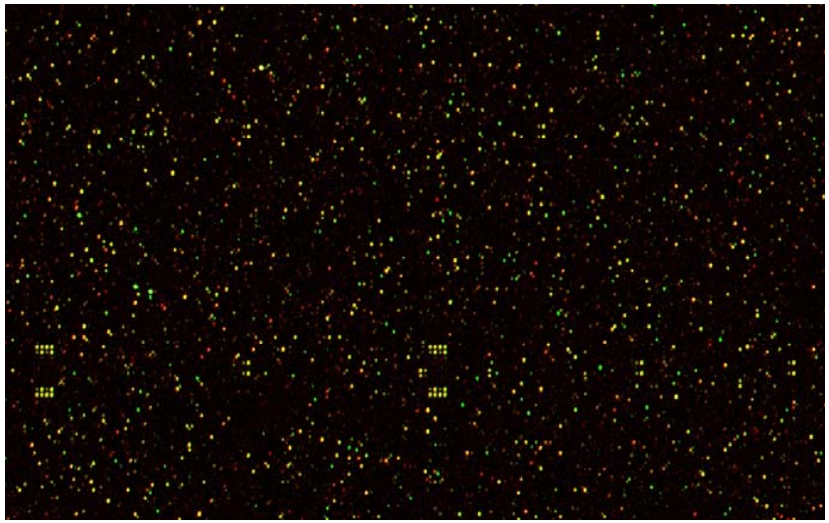
1) Mapping Chromatin Modifications

2) Mapping DNA Methylation

Mapping Transcribed Regions



- Probe Affymetrix or Nimblegen Arrays with cDNA
- >50% of transcribed regions not annotated



ENCODE Project: Transcribed and Regulatory Regions

Future Directions

1) Whole genome sequencing.

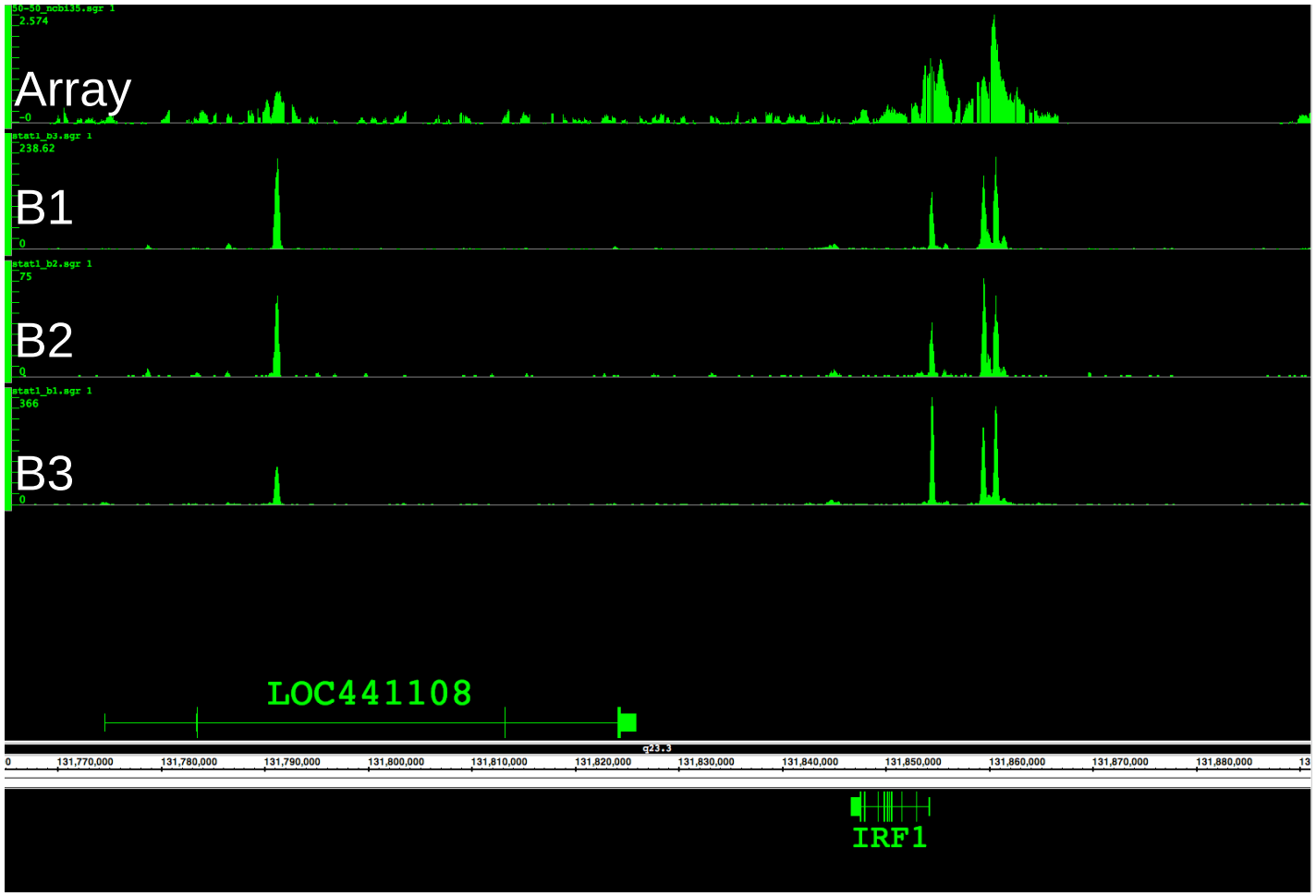
2) Whole genome mapping of transcribed regions and regulatory information in small samples.

3) Proteomics approaches.

Technologies for Large Population Studies: Possible Recommendations

1) High density SNPs
2) High resolution CNVs/SVs 2 kb or larger
3) Whole genome gene expression (coding and noncoding). What tissues?
4) Sera proteome and metabolomics analysis. Other fluids?
5) Whole genome TF/Methylation mapping. Which ones? What tissues?

STAT1 x 3 Reps & Array (IRF1)



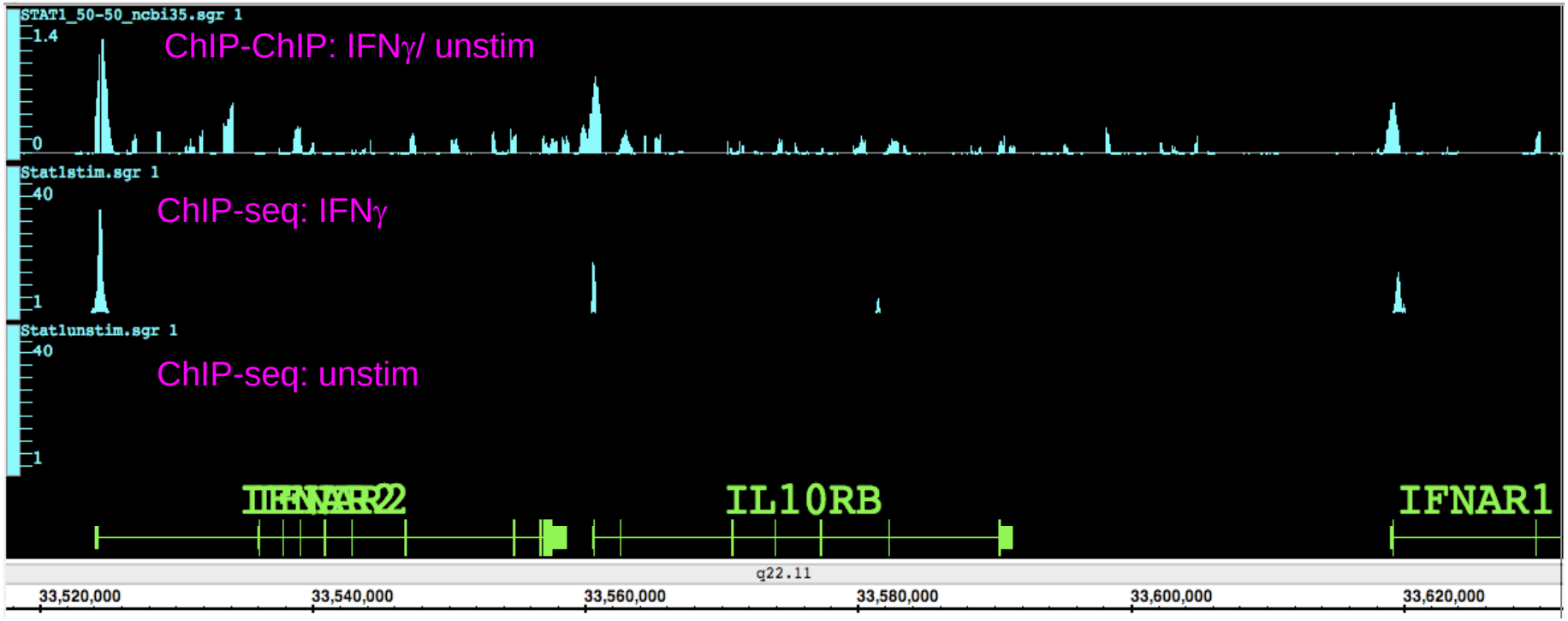
Mapped > 40K STAT1 Binding Sites

STAT1 ChIP-Seq on IFN Treated HeLa Cells using Illumina Sequencing

with the University of British Columbia

ChIP-chip (NimbleGen 2.1 M feature arrays)

ChIP-seq (**15.1 M** mapped reads for the IFNγ)



Robertson et al. 2007

ENCODE Pilot Phase
(Analyze 1% of the Human Genome)

Mapping transcribed regions

13 Experiments

Mapping regulatory information

>150 ChIP Chip experiments

DNAase hypersensitive sites

Massively Parallel Sequencing Technologies

454

400K reads

250 b

Illumina and ABI

>40M reads

36 b

Distribution of SVs