Uncovering hidden genes in intergenic GWAS regions

Mike Clark
Cancer Genomics Research Laboratory
Garvan Institute for Medical Research
m.clark@garvan.org.au

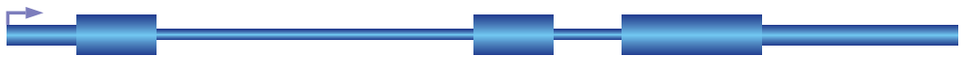
The University Of Queensland AUSTRALIA | IMB Institute for Molecular Bioscience

GARVAN INSTITUTE
Breakthrough Medical Research

Uncovering hidden genes in intergenic GWAS regions with RNA capture sequencing

- The pervasive transcriptome and “intergenic” genome-wide association study (GWAS) regions
- Detecting transcription with RNA Capture sequencing
- Using capture sequencing for novel gene discovery in human “intergenic” GWAS regions

Transcription is complex, and pervasive across the genome



noncoding RNA

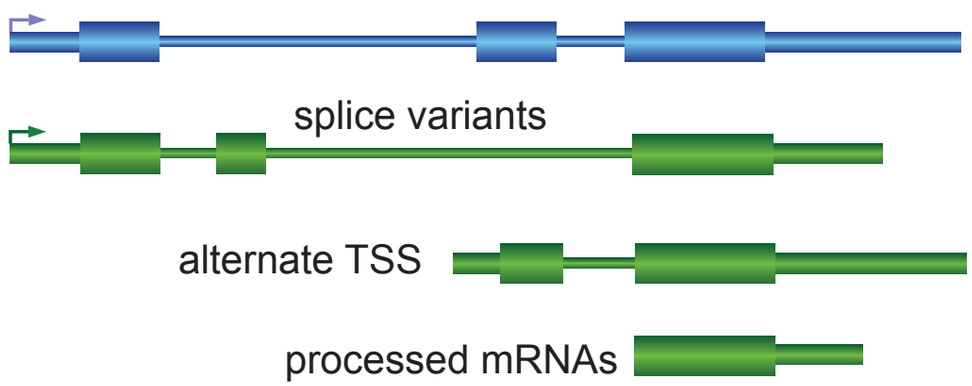
coding gene

coding gene variants

Sense strand

AntiSense strand

Transcription is complex, and pervasive across the genome



noncoding RNA

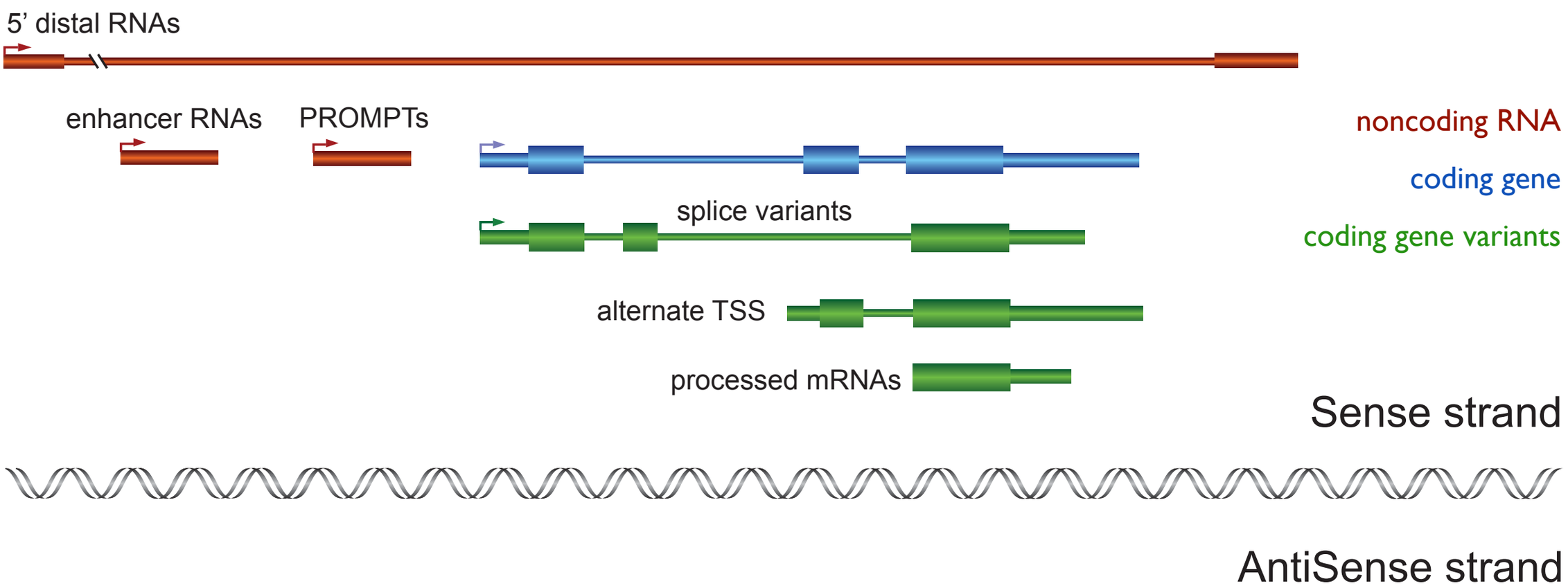
coding gene

coding gene variants

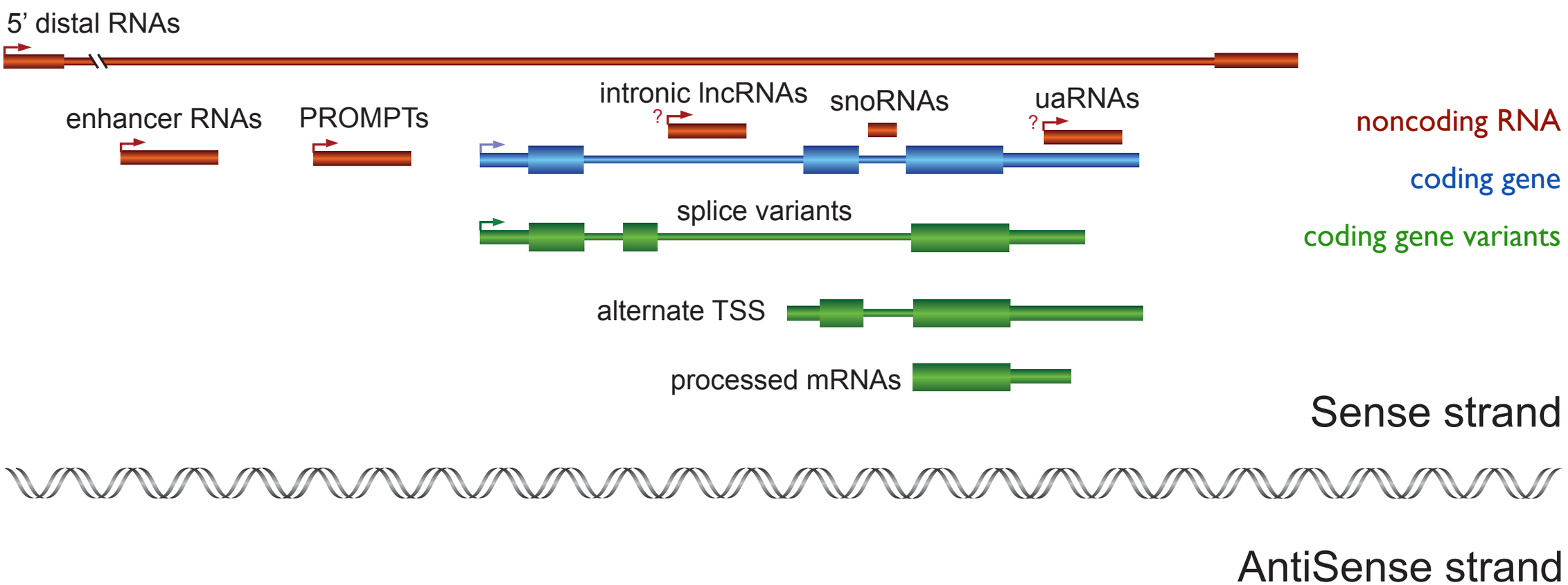
Sense strand

AntiSense strand

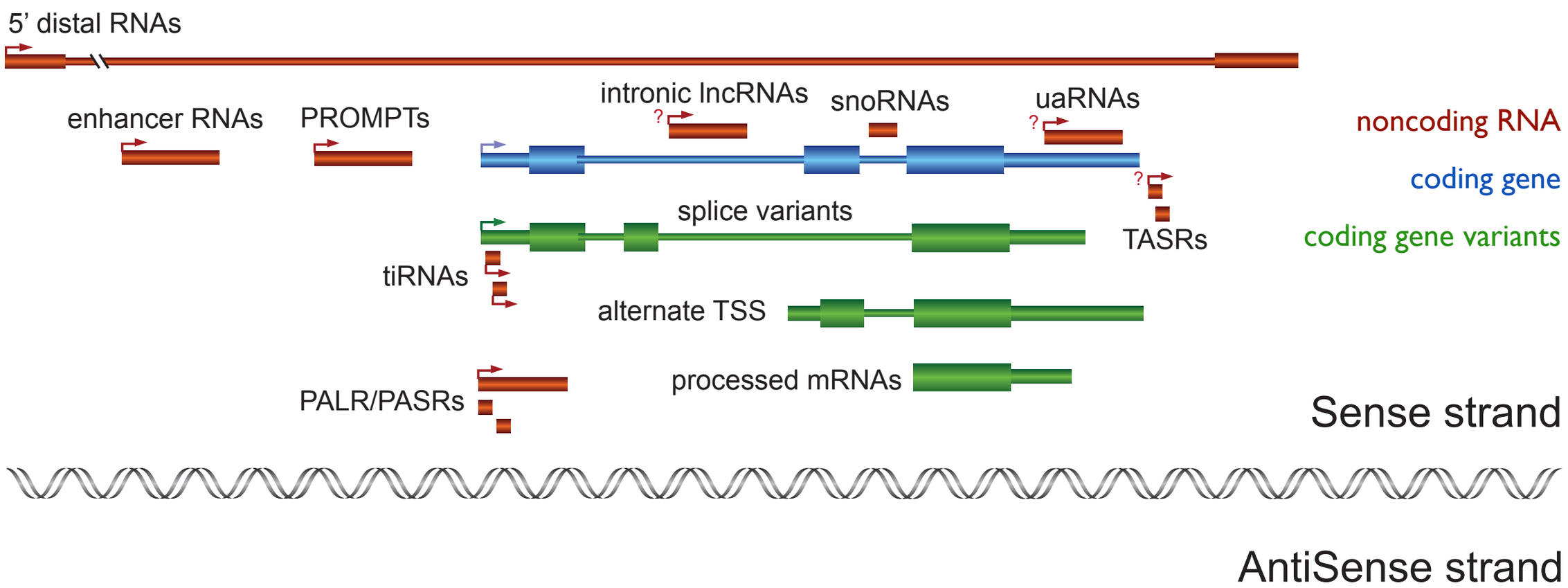
Transcription is complex, and pervasive across the genome



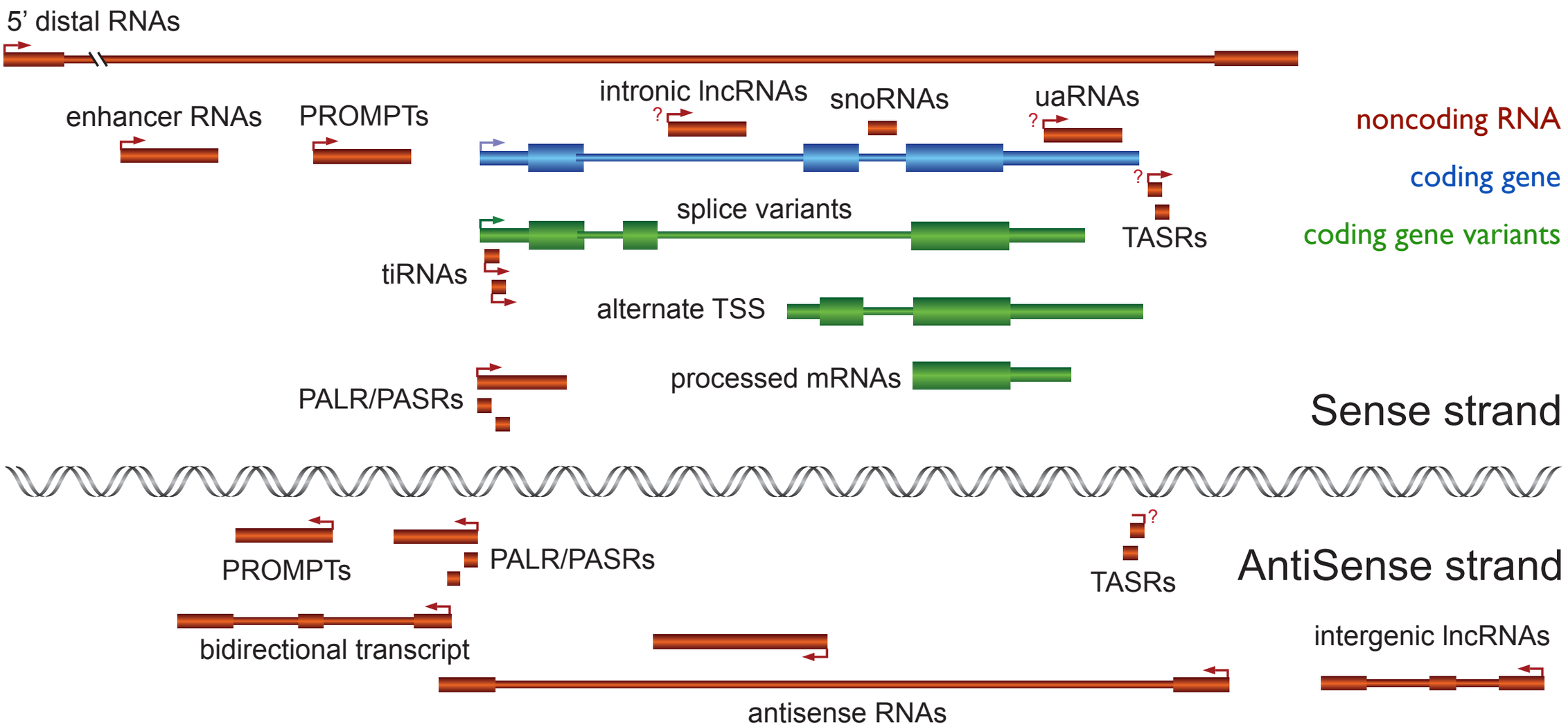
Transcription is complex, and pervasive across the genome



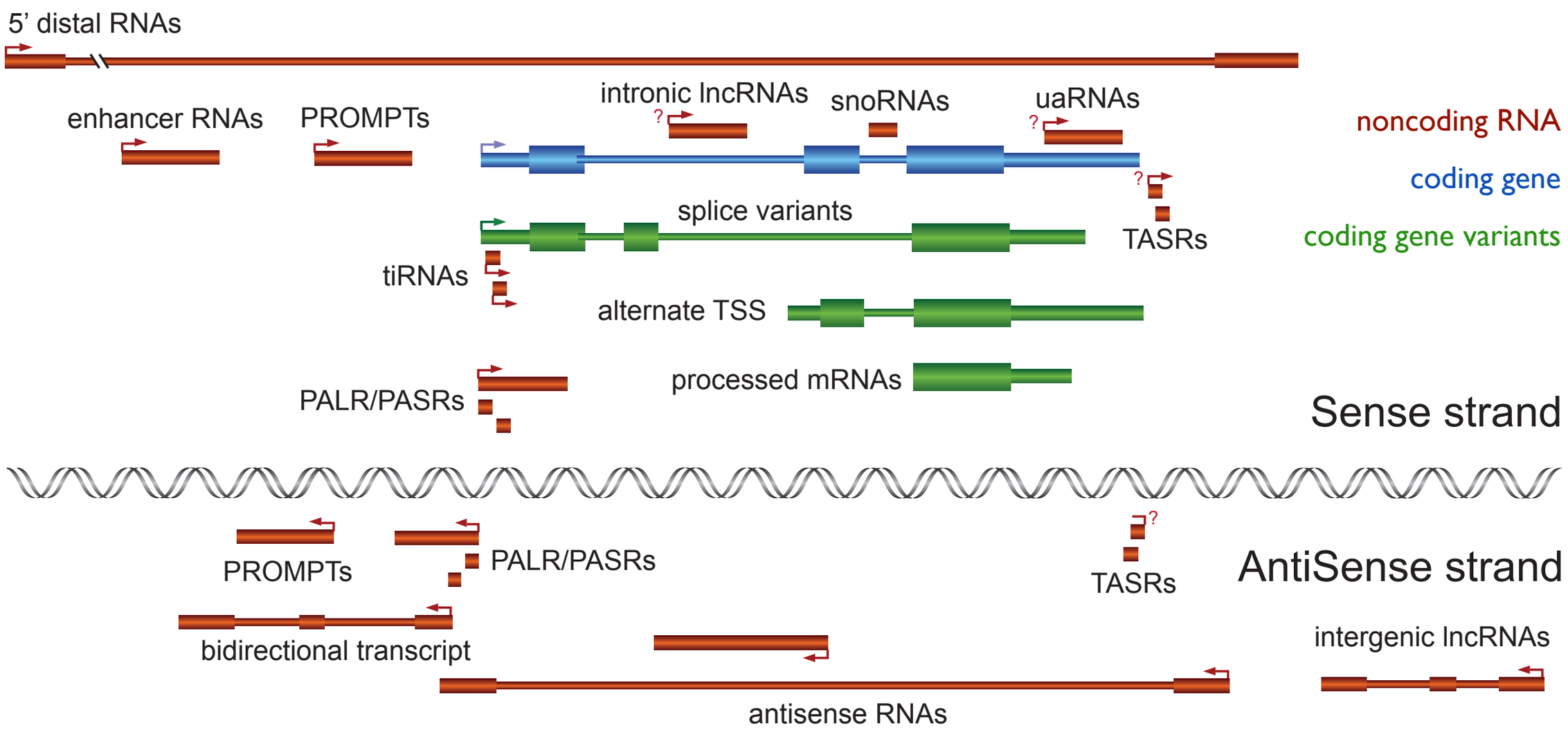
Transcription is complex, and pervasive across the genome



Transcription is complex, and pervasive across the genome



Transcription is complex, and pervasive across the genome



- At least 75% of the human genome is transcribed, though a lot of this is transcribed at low levels (ENCODE 2012)

Transcripts can be lowly and specifically expressed

High expression

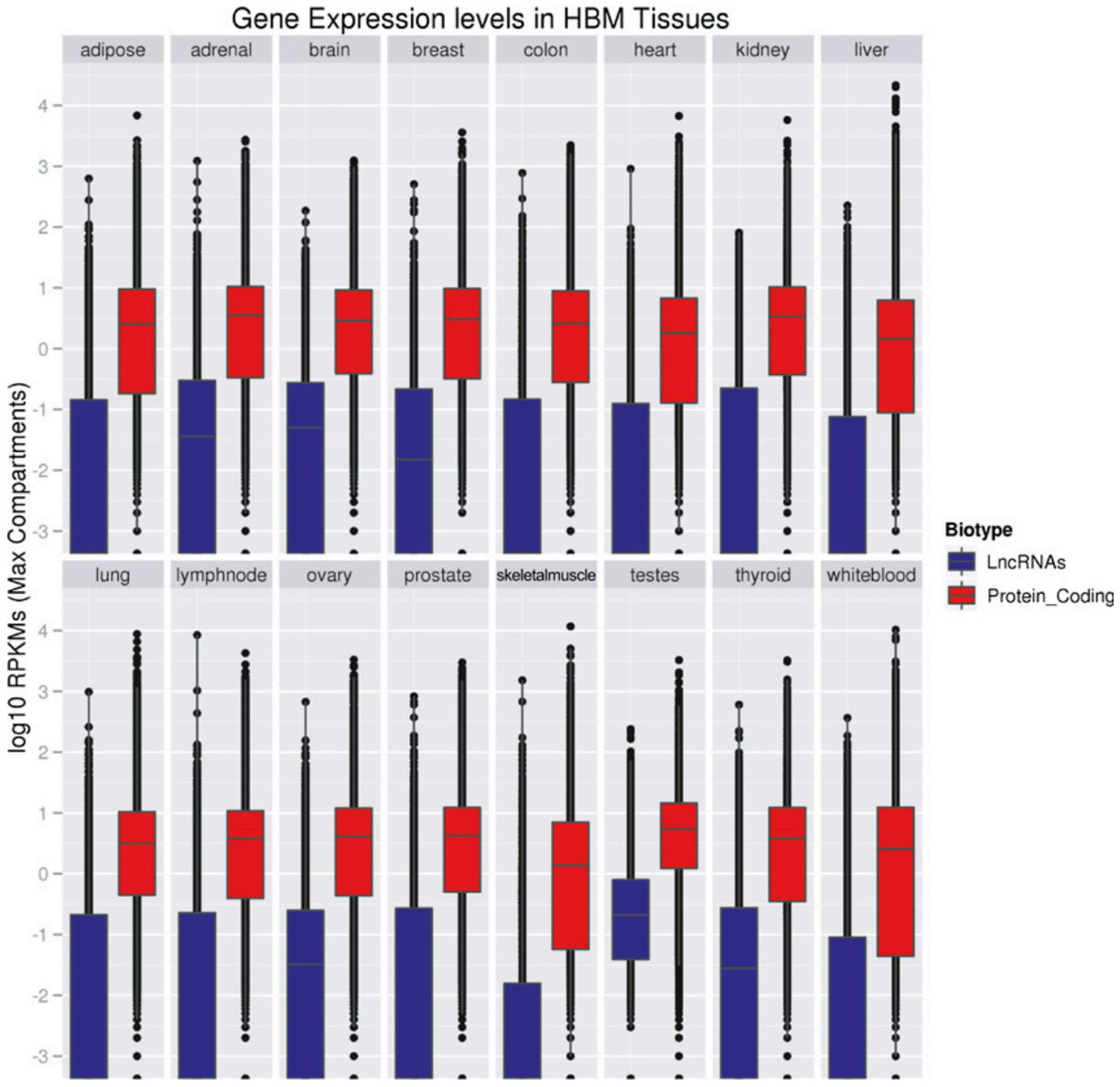
Low expression

High expression

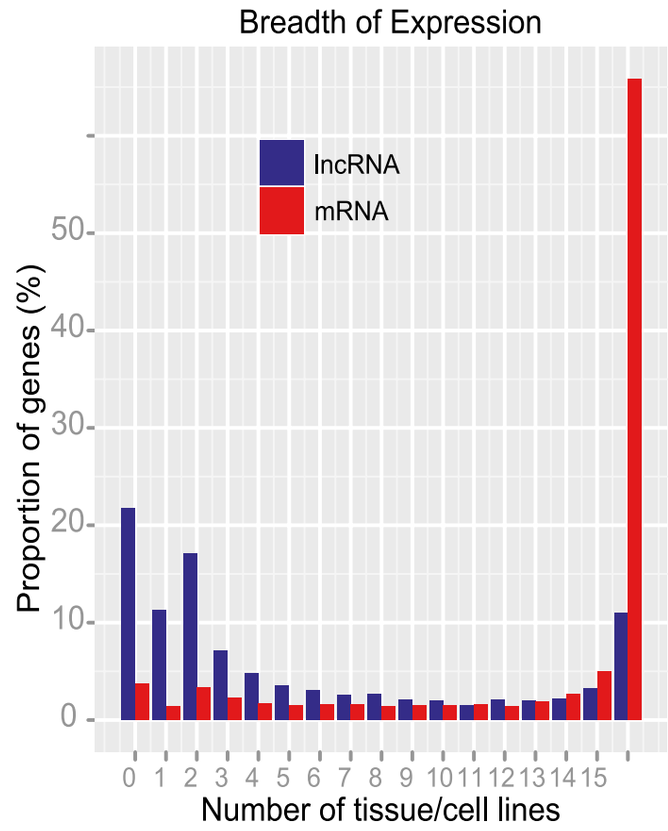
Low expression

A

Different tissues



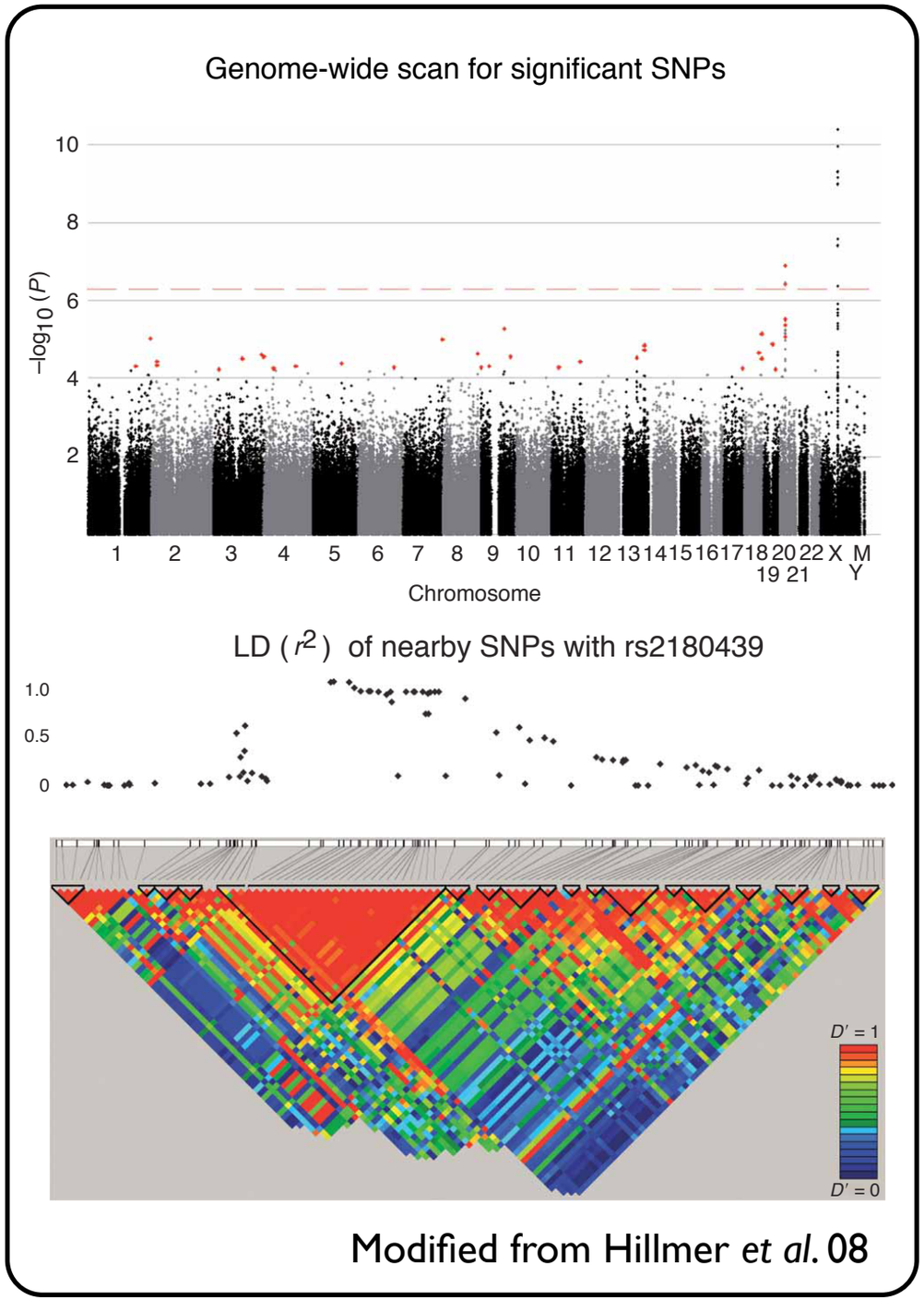
B



Red: coding genes. Blue: lncRNAs

ENCODE 2012

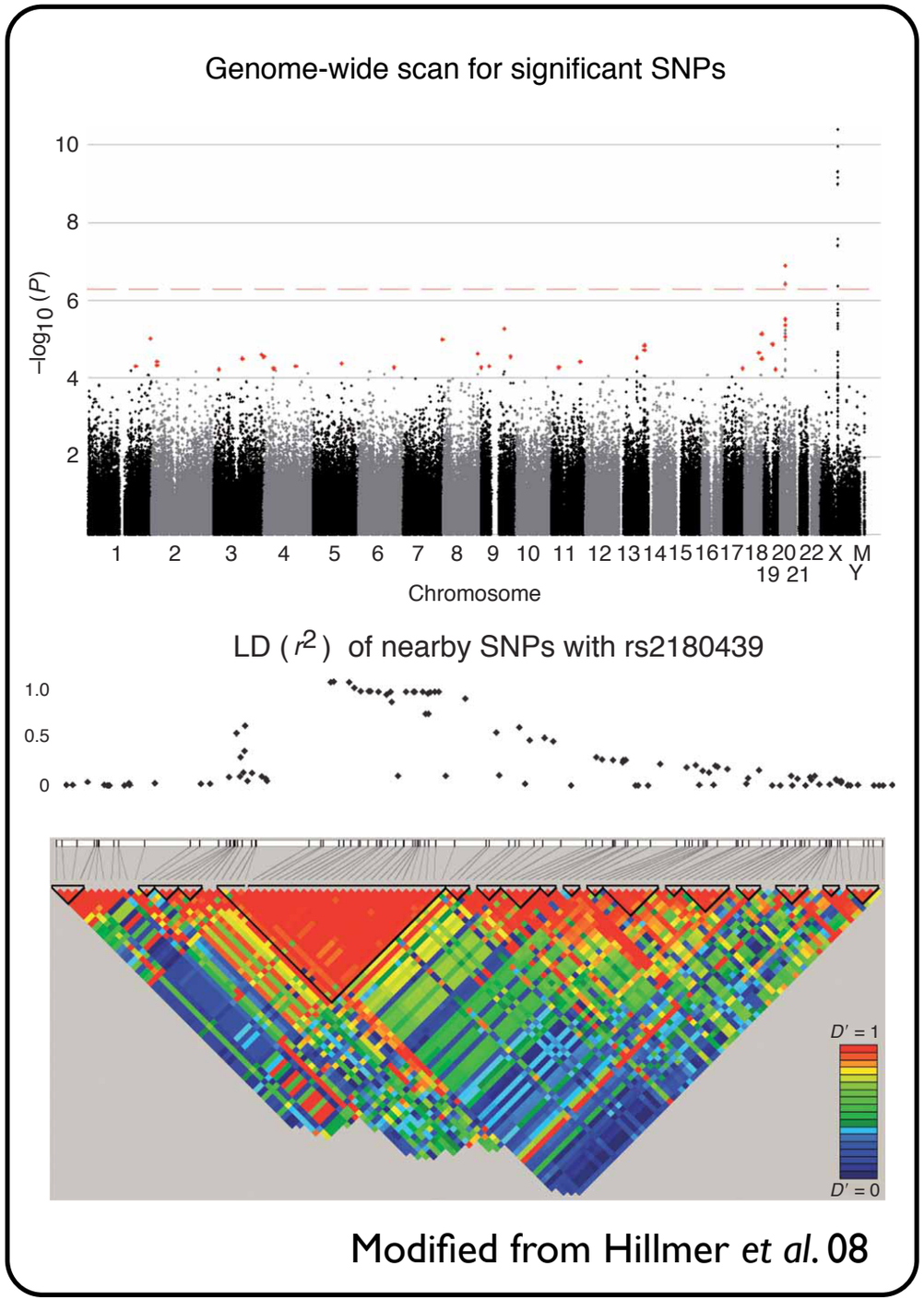
“Intergenic” GWAS regions



Modified from Hillmer *et al.* 08

“Intergenic” GWAS regions

- Genome-wide association studies identify SNPs with a significant association to a disease or trait

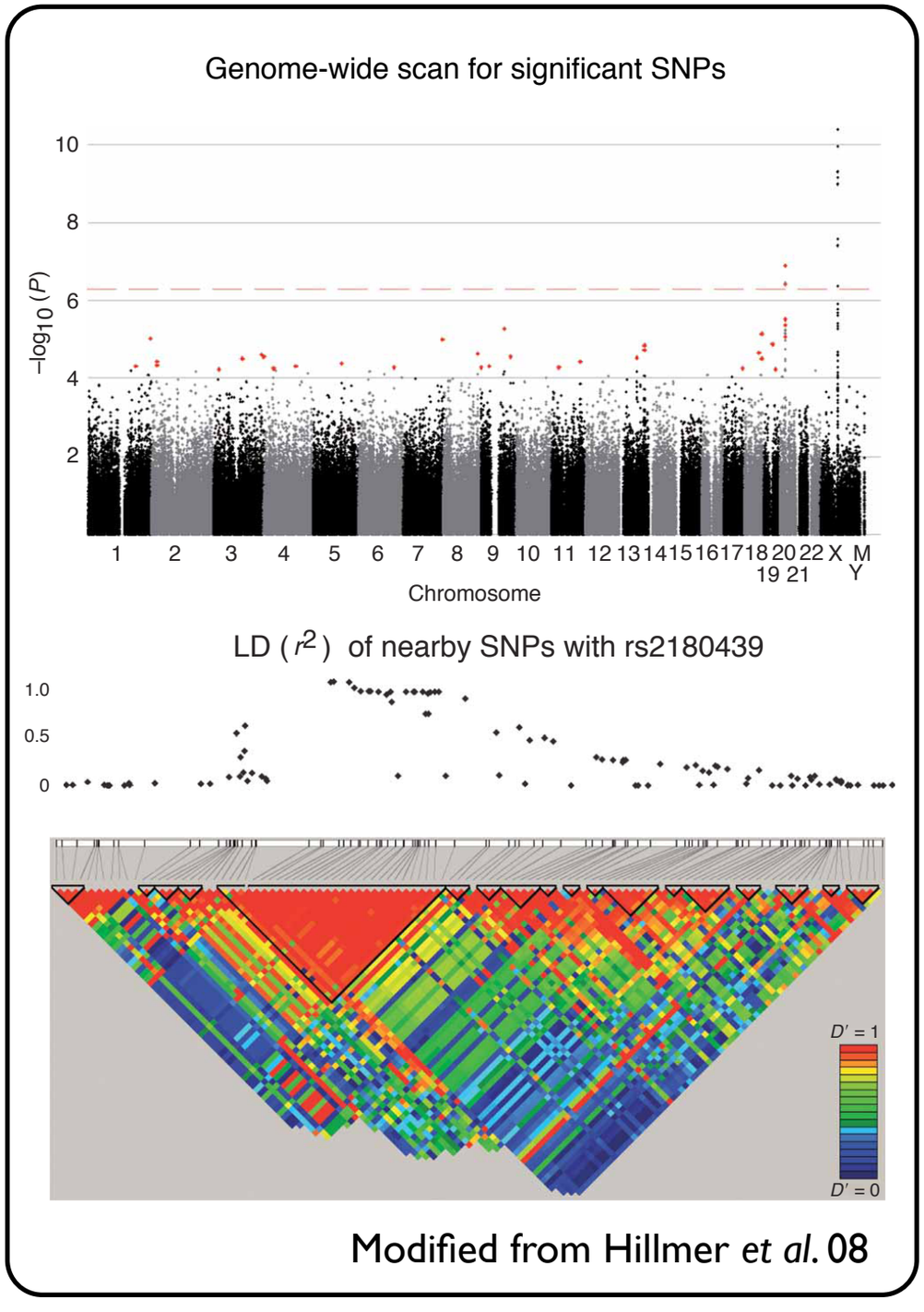


Modified from Hillmer *et al.* 08

“Intergenic” GWAS regions

- Genome-wide association studies identify SNPs with a significant association to a disease or trait

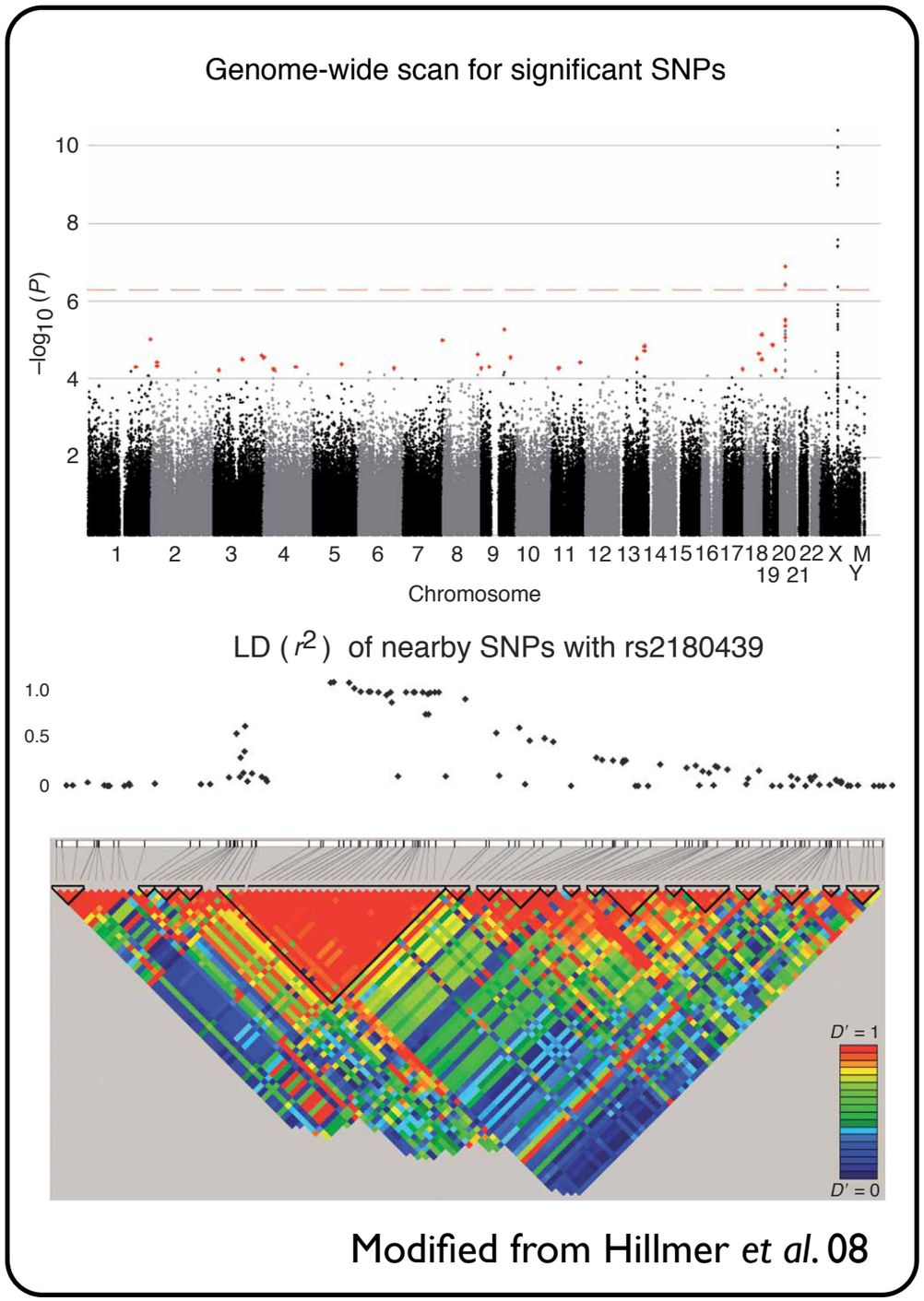
- 88% of identified SNPs are intergenic or intronic (Hindorff *et al.* 09)



Modified from Hillmer *et al.* 08

“Intergenic” GWAS regions

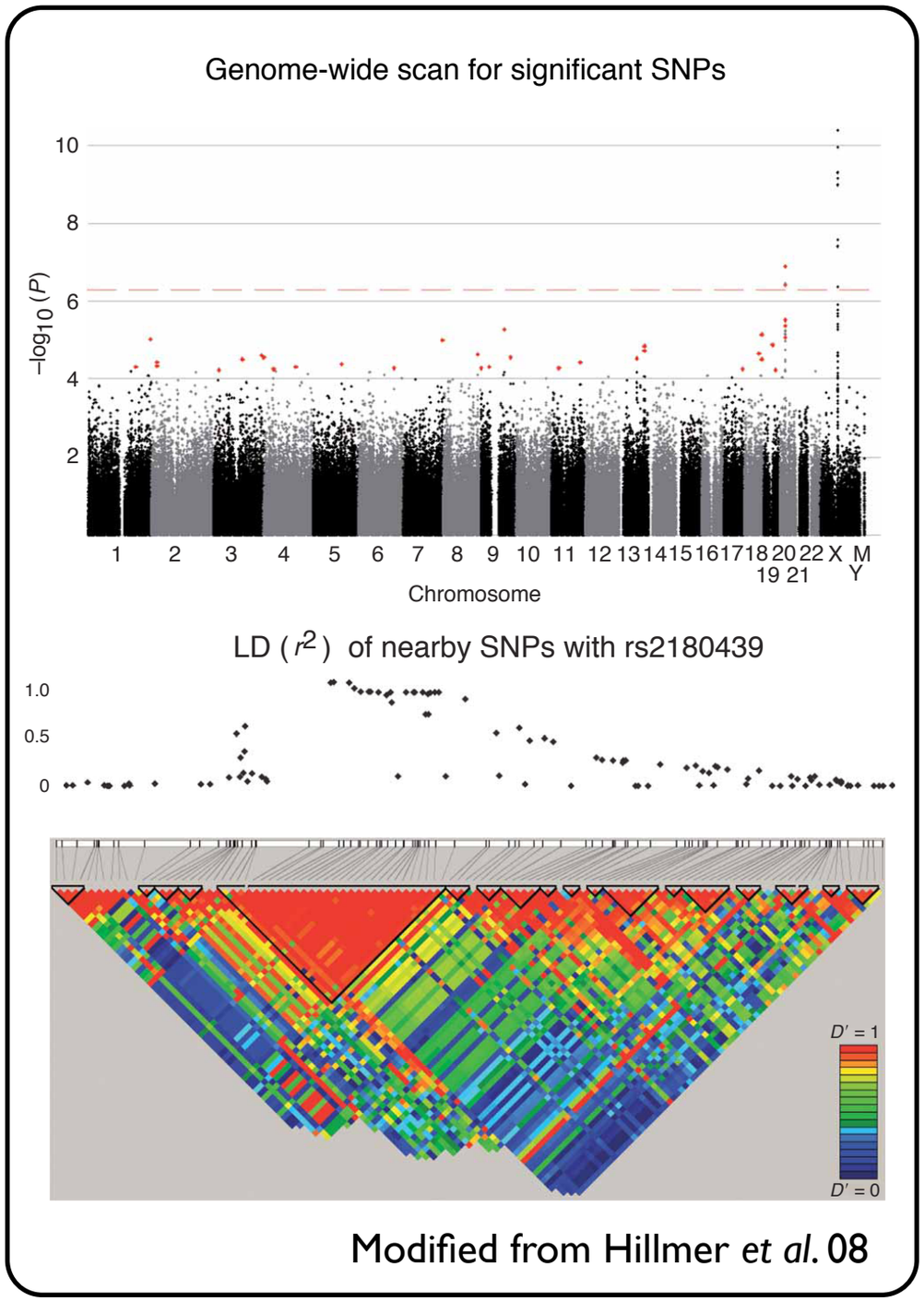
- Genome-wide association studies identify SNPs with a significant association to a disease or trait
- 88% of identified SNPs are intergenic or intronic (Hindorff *et al.* 09)
- SNP may not be causative allele, but this should be in a region of linkage disequilibrium (LD) around SNP



Modified from Hillmer *et al.* 08

“Intergenic” GWAS regions

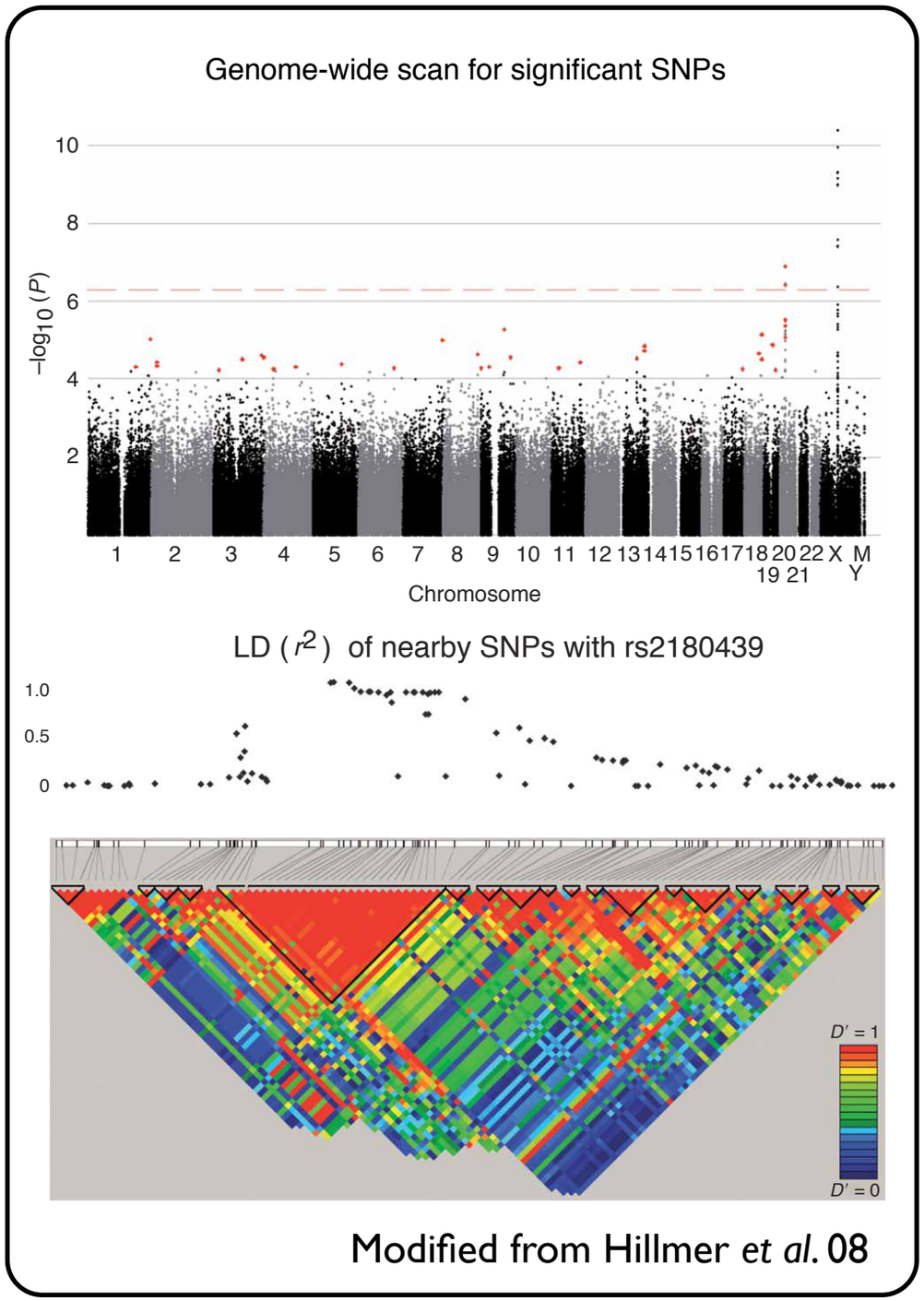
- Genome-wide association studies identify SNPs with a significant association to a disease or trait
- 88% of identified SNPs are intergenic or intronic (Hindorff *et al.* 09)
- SNP may not be causative allele, but this should be in a region of linkage disequilibrium (LD) around SNP
- Hundreds of GWAS LD regions are intergenic.



Modified from Hillmer *et al.* 08

“Intergenic” GWAS regions

- Genome-wide association studies identify SNPs with a significant association to a disease or trait
- 88% of identified SNPs are intergenic or intronic (Hindorff *et al.* 09)
- SNP may not be causative allele, but this should be in a region of linkage disequilibrium (LD) around SNP
- Hundreds of GWAS LD regions are intergenic.
- Are these really empty of genes, or have we just not found them yet?



Modified from Hillmer *et al.* 08

“Intergenic” GWAS regions

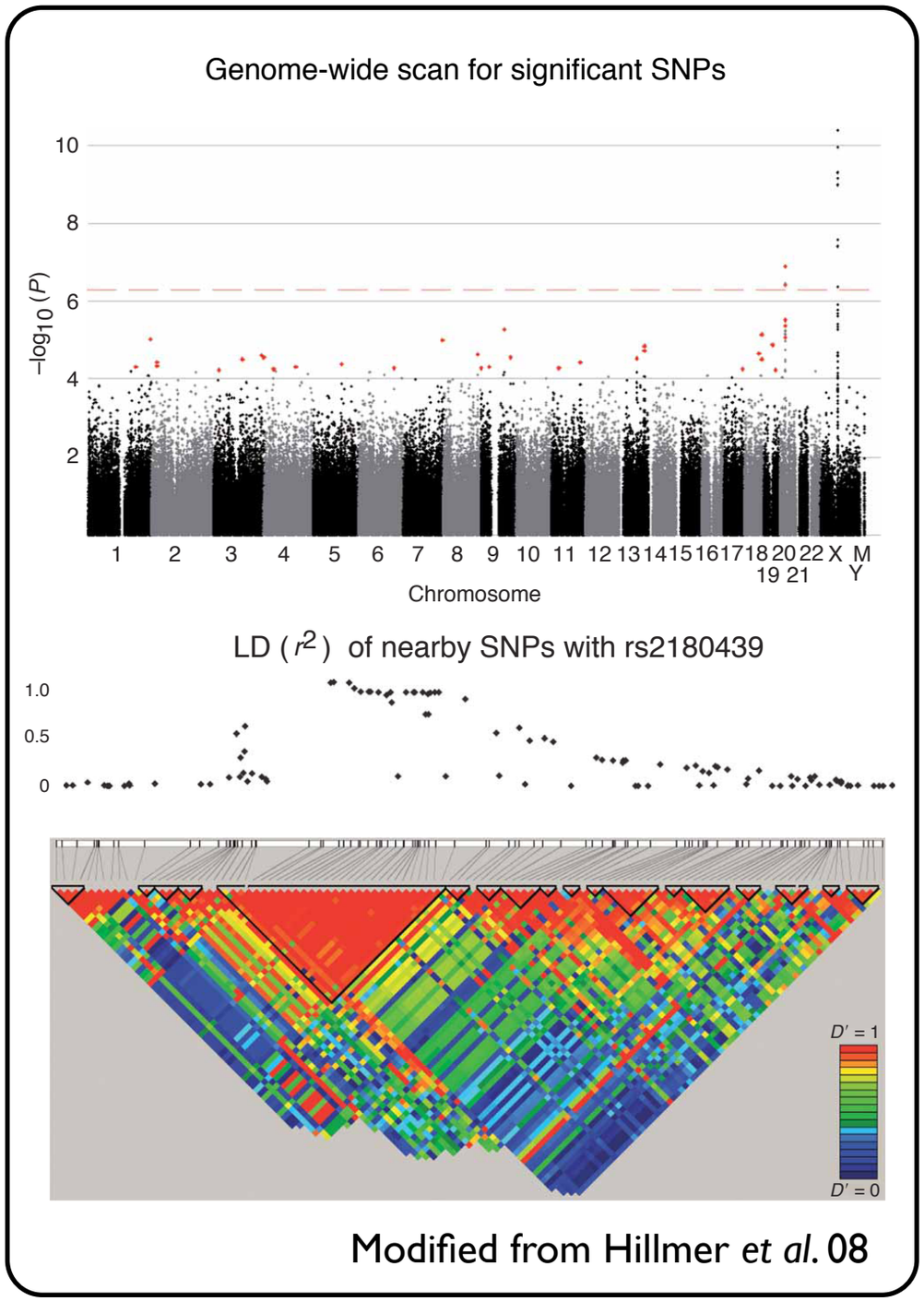
- Genome-wide association studies identify SNPs with a significant association to a disease or trait

- 88% of identified SNPs are intergenic or intronic (Hindorff *et al.* 09)

- SNP may not be causative allele, but this should be in a region of linkage disequilibrium (LD) around SNP

- Hundreds of GWAS LD regions are intergenic.

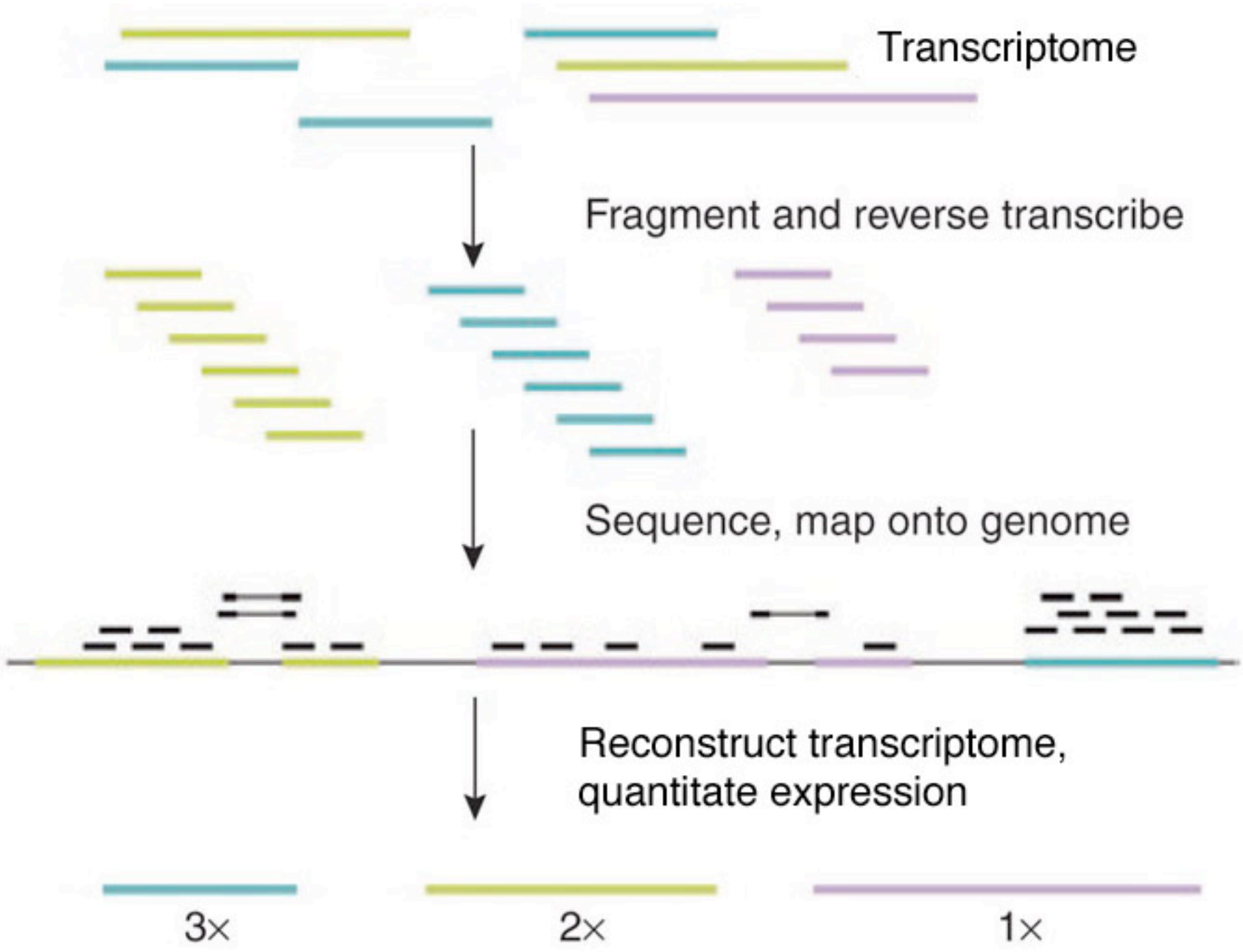
- Are these really empty of genes, or have we just not found them yet?



Modified from Hillmer *et al.* 08

- Aim to discover novel genes within these intergenic GWAS regions

Standard RNA sequencing outline



Modified from Pepke 09

RNAseq is mostly re-sequencing of highly expressed genes

The human transcriptome is complex. RNA Sequencing is like eating some of these jelly beans at random. You have a good chance of picking common colours but a low chance of finding rare ones.

RNAseq is mostly re-sequencing of highly expressed genes

The human transcriptome is complex. RNA Sequencing is like eating some of these jelly beans at random. You have a good chance of picking common colours but a low chance of finding rare ones.

RNAseq is mostly re-sequencing of highly expressed genes

The human transcriptome is complex. RNA Sequencing is like eating some of these jelly beans at random. You have a good chance of picking common colours but a low chance of finding rare ones.

Limitations of RNAseq: Rare or restricted transcripts

Limitations of RNAseq: Rare or restricted transcripts

- Rare transcripts:

  The most abundant 1.5% of transcripts take up almost half the RNAseq reads, while only 1% of reads are measuring the least abundant 44% of detected transcripts (Jiang *et al.* 11).

Limitations of RNAseq: Rare or restricted transcripts

- Rare transcripts:

  The most abundant 1.5% of transcripts take up almost half the RNAseq reads, while only 1% of reads are measuring the least abundant 44% of detected transcripts (Jiang *et al*. 11).

- Transcripts with restricted expression:

  A transcript expressed at an average level (10 copies per cell) in 0.1% of brain cells (~170 million cells), requires ~400 million reads for 1x coverage.

  8x coverage is required for robust assembly (Jiang *et al*. 11)..

Limitations of RNAseq: Rare or restricted transcripts

- Rare transcripts:

  The most abundant 1.5% of transcripts take up almost half the RNAseq reads, while only 1% of reads are measuring the least abundant 44% of detected transcripts (Jiang *et al*. 11).

- Transcripts with restricted expression:

  A transcript expressed at an average level (10 copies per cell) in 0.1% of brain cells (~170 million cells), requires ~400 million reads for 1x coverage.

  8x coverage is required for robust assembly (Jiang *et al*. 11)..

- So how do we find the rare or restricted transcripts likely present in intergenic GWAS regions?

Limitations of RNAseq: Rare or restricted transcripts

- Rare transcripts:

  The most abundant 1.5% of transcripts take up almost half the RNAseq reads, while only 1% of reads are measuring the least abundant 44% of detected transcripts (Jiang *et al*. 11).

- Transcripts with restricted expression:

  A transcript expressed at an average level (10 copies per cell) in 0.1% of brain cells (~170 million cells), requires ~400 million reads for 1x coverage.

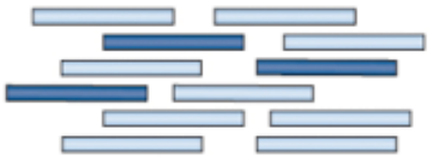
  8x coverage is required for robust assembly (Jiang *et al*. 11)..

- RNA Capture Sequencing provides a potential method to detect transcripts with rare or restricted expression.

RNA CaptureSeq

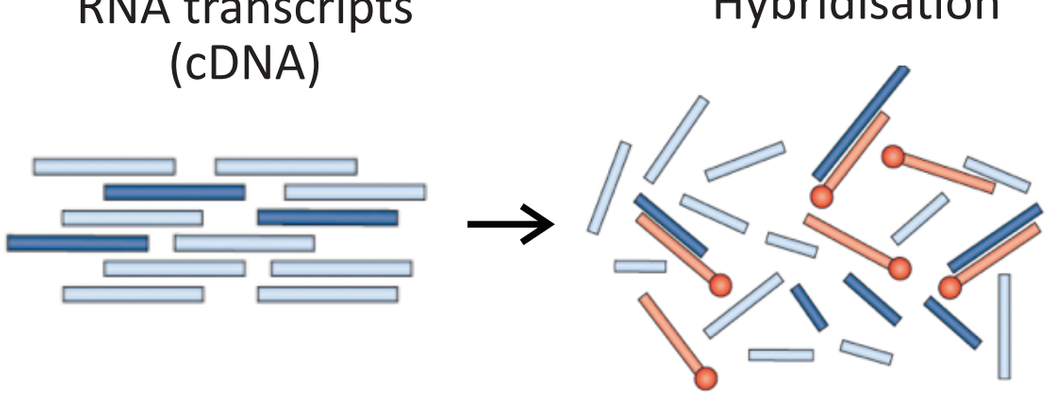
CaptureSeq selects a portion of the transcriptome for focused sequencing, thereby achieving a huge increase in sequencing depth and coverage.

RNA transcripts
(cDNA)



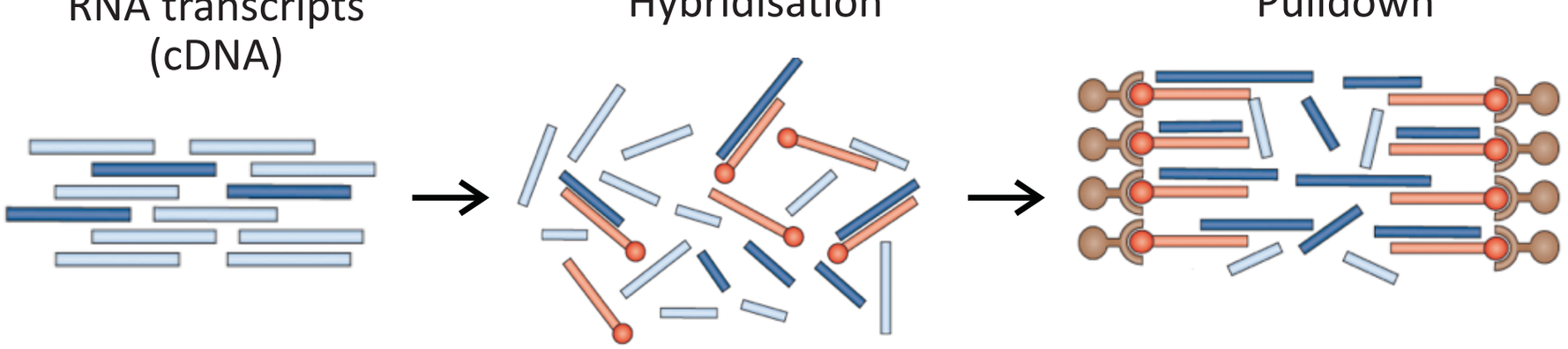
RNA CaptureSeq

CaptureSeq selects a portion of the transcriptome for focused sequencing, thereby achieving a huge increase in sequencing depth and coverage.



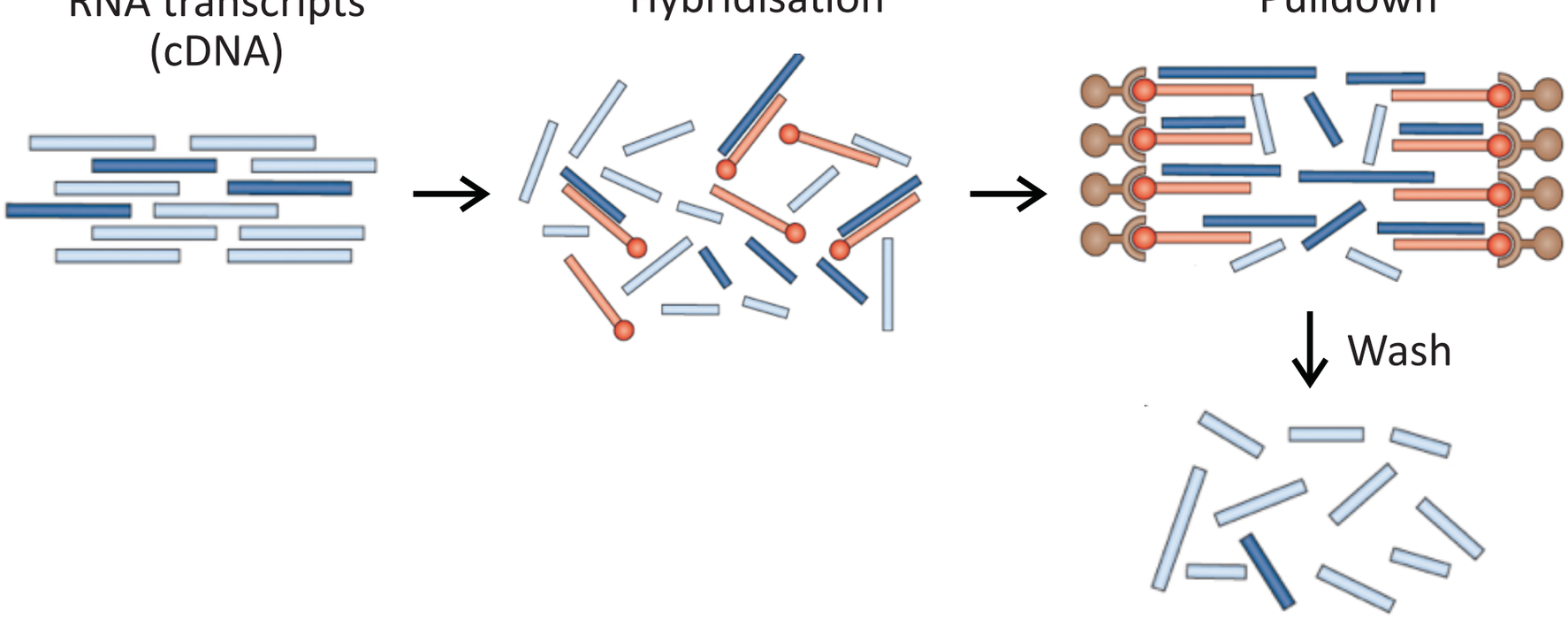
RNA CaptureSeq

CaptureSeq selects a portion of the transcriptome for focused sequencing, thereby achieving a huge increase in sequencing depth and coverage.



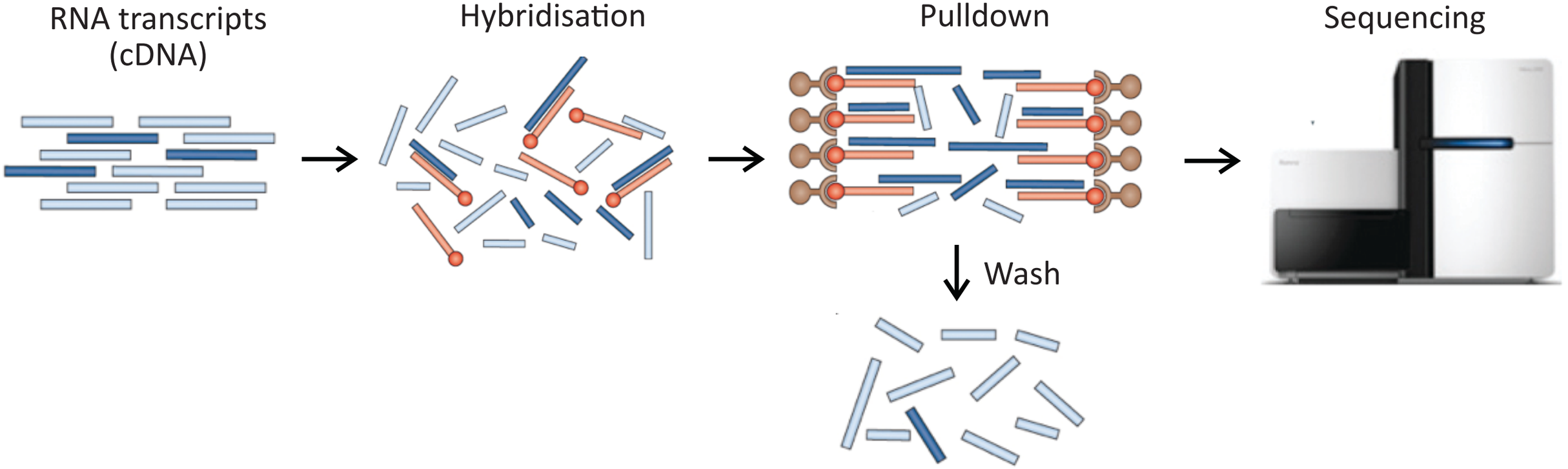
RNA CaptureSeq

CaptureSeq selects a portion of the transcriptome for focused sequencing, thereby achieving a huge increase in sequencing depth and coverage.



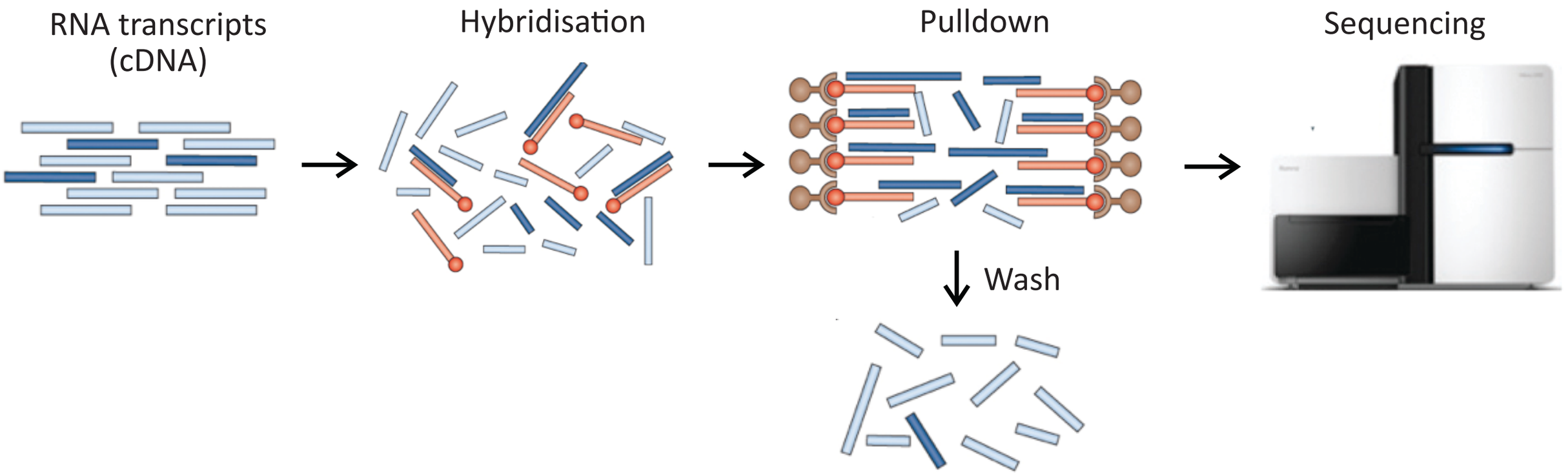
RNA CaptureSeq

CaptureSeq selects a portion of the transcriptome for focused sequencing, thereby achieving a huge increase in sequencing depth and coverage.

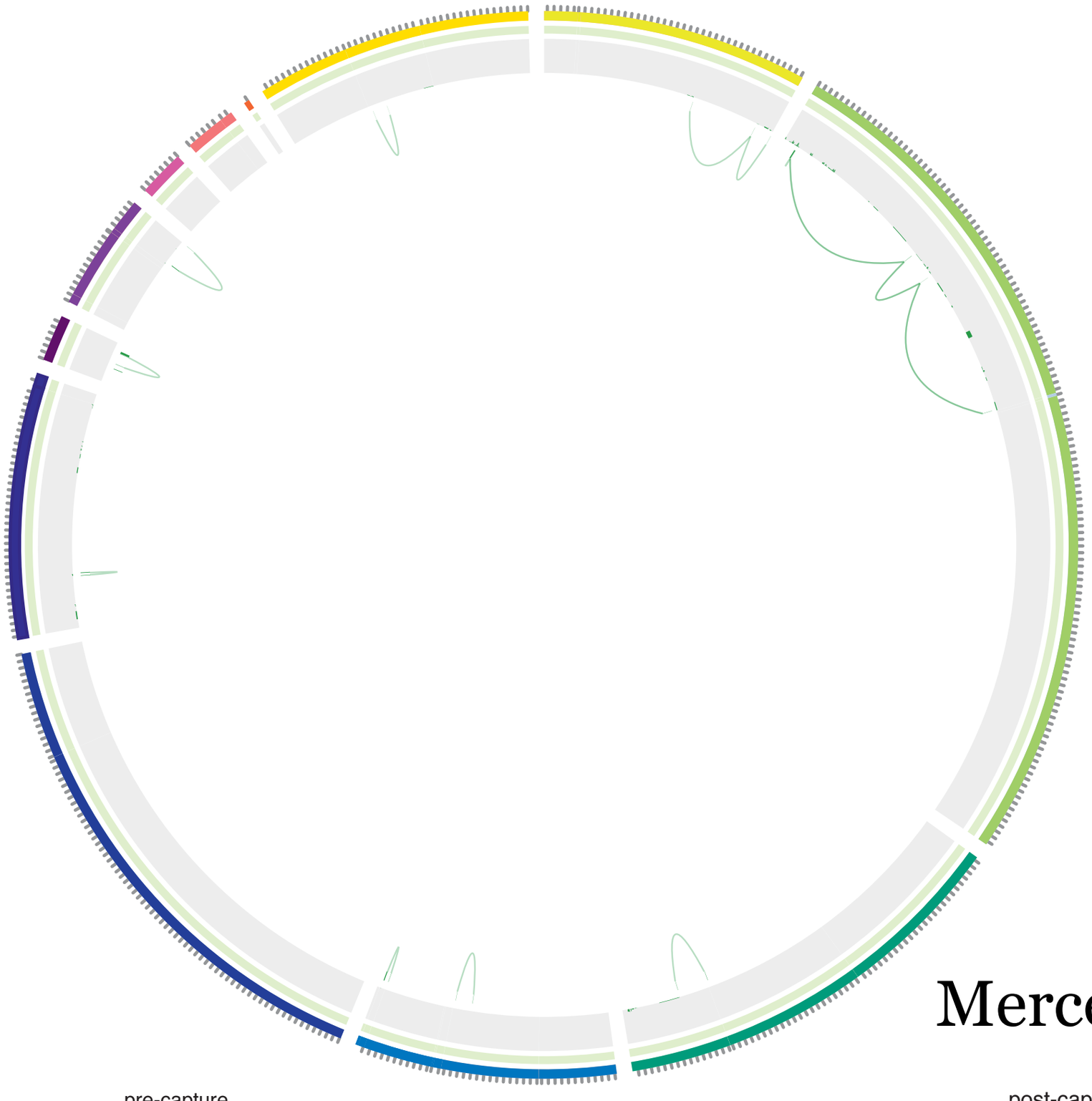


RNA CaptureSeq

This approach can characterise transcripts with rare or transient expression that is below the detection limits of conventional sequencing approaches.



RNA capture sequencing can discover new transcripts



Mercer et al. 2011

Tracks
(from outer edge):

- Annotated gene
- pre-capture
  - RNAseq
  - Assembled exon
  - Splice junction
- Probed region
- post-capture
  - RNAseq
  - Novel assembled exon
  - Novel splice junction

RNA capture sequencing can discover new transcripts

Mercer et al. 2011

Tracks
(from outer edge):

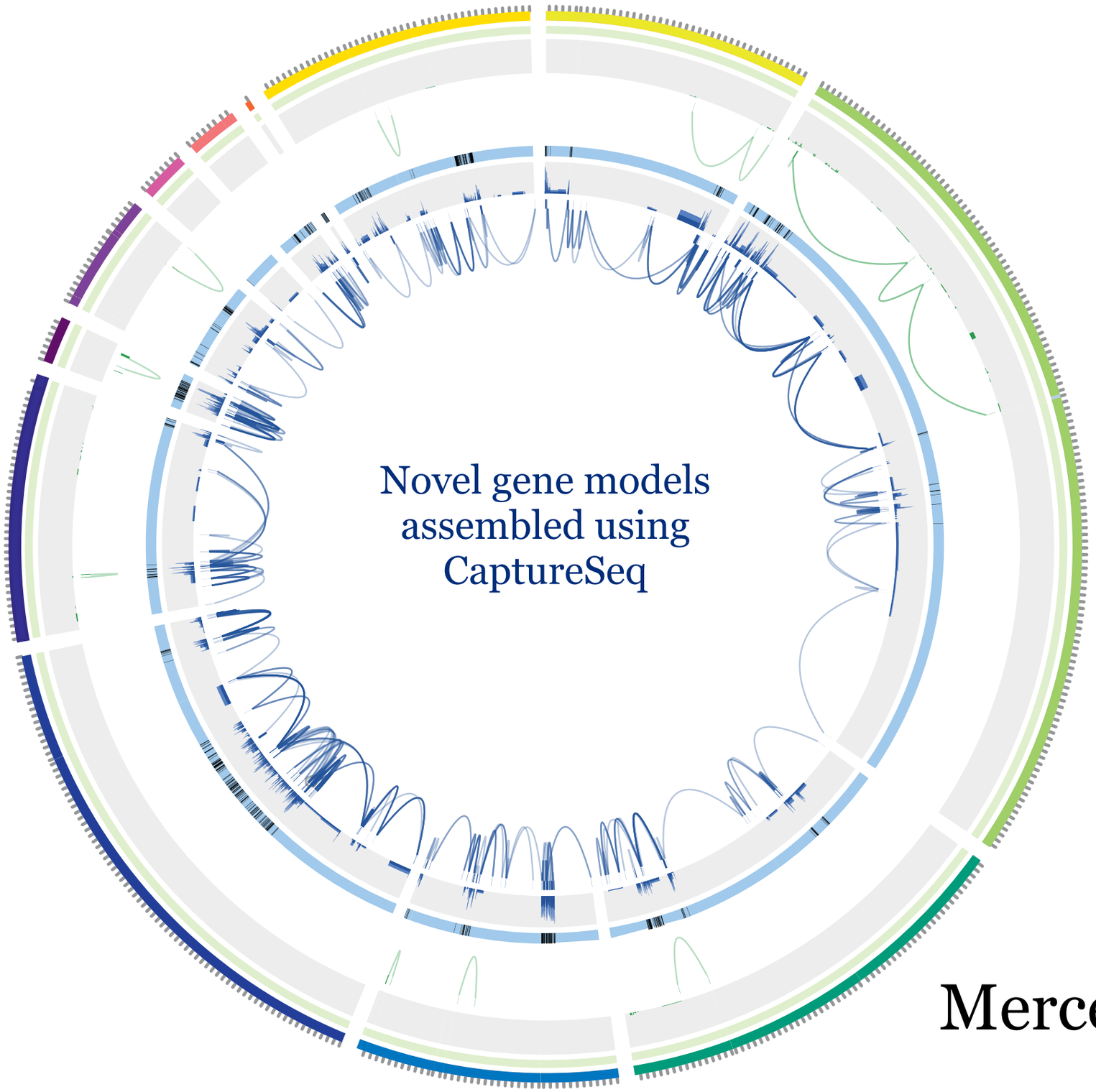
Annotated gene

pre-capture: RNAseq, Assembled exon, Splice junction

Probed region

post-capture: RNAseq, Novel assembled exon, Novel splice junction

RNA capture sequencing can discover new transcripts



Mercer et al. 2011

Tracks
(from outer edge):

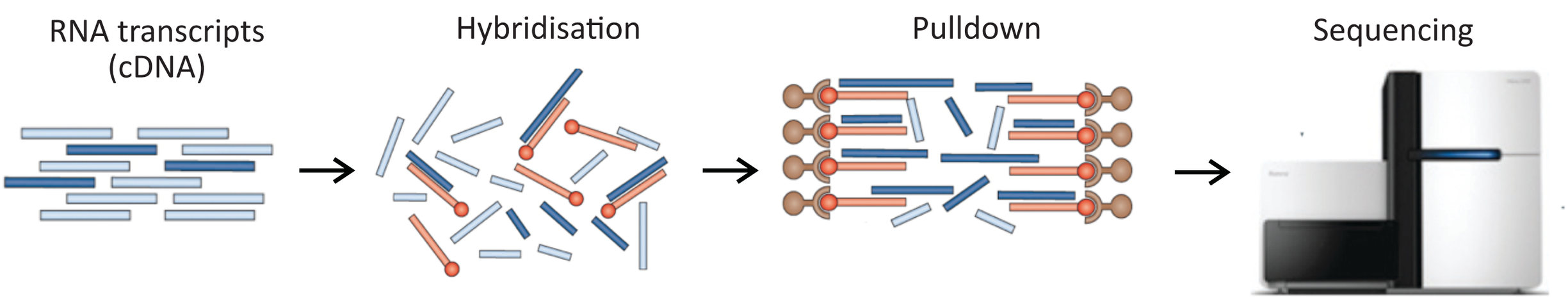
Annotated gene

pre-capture: RNAseq, Assembled exon, Splice junction

Probed region

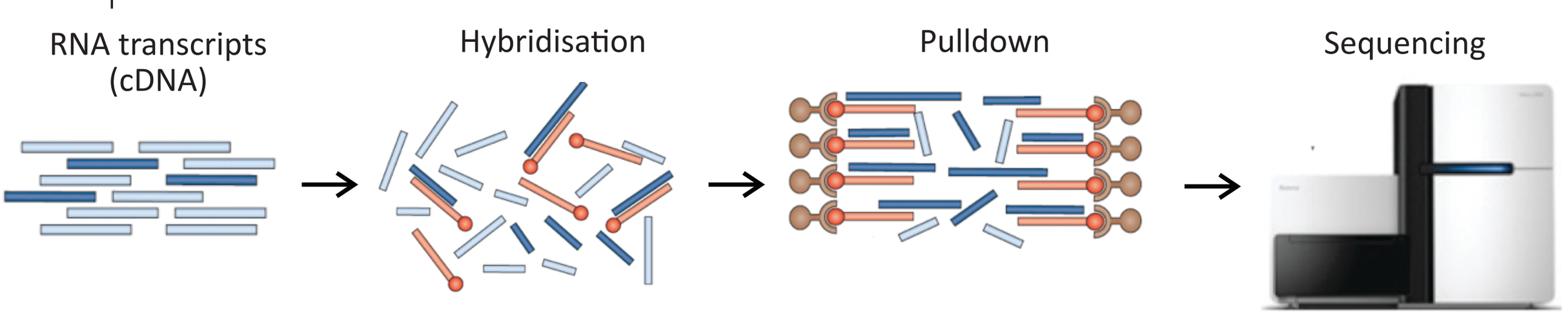
post-capture: RNAseq, Novel assembled exon, Novel splice junction

Identifying novel genes in “intergenic” GWAS regions



Identifying novel genes in “intergenic” GWAS regions

Ribo-depleted RNA from 20 normal human tissues.



Identifying novel genes in “intergenic” GWAS regions

Ribo-depleted RNA from 20 normal human tissues.

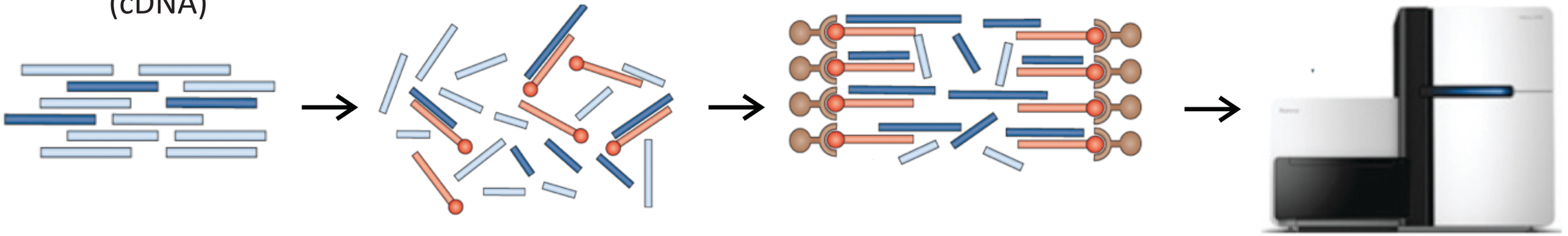
Probes to 339 regions in linkage disequilibrium with GWAS SNP that contained no genes.

RNA transcripts (cDNA)

Hybridisation

Pulldown

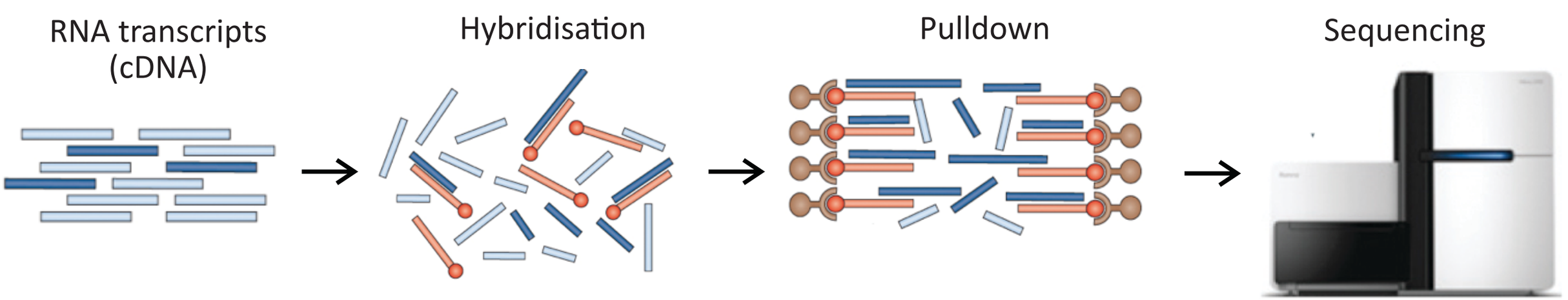
Sequencing



Identifying novel genes in “intergenic” GWAS regions

Ribo-depleted RNA from 20 normal human tissues.

Probes to 339 regions in linkage disequilibrium with GWAS SNP that contained no genes.



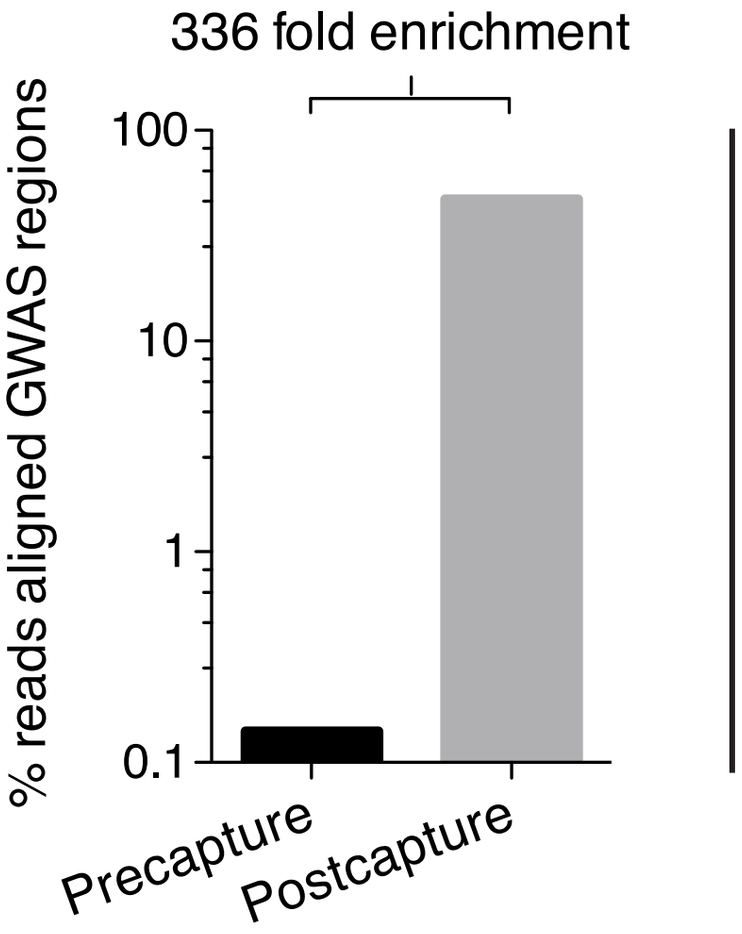
Advantage of GWAS regions over other intergenic regions is we know these regions are functionally relevant.

Targeting “intergenic” GWAS regions for ~150 traits and diseases

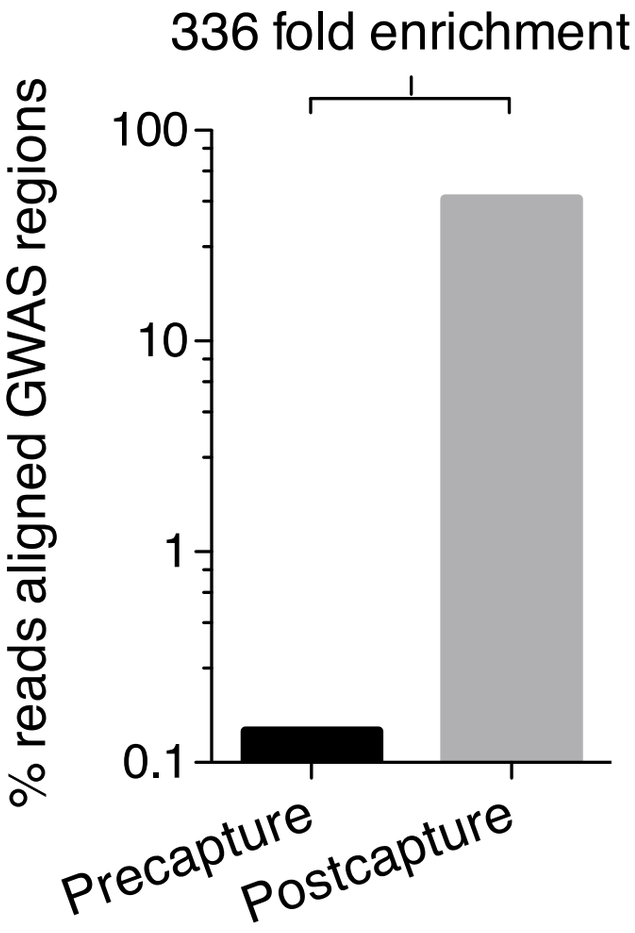
AB1-42
Adiponectin levels
Age-related macular degeneration
Aging (time to event)
Aging traits
AIDS
Alcoholism
Allergic rhinitis
Alzheimer's disease
Amyotrophic lateral sclerosis
Ankylosing spondylitis
Anthropometric traits
Aortic root size
Aortic stiffness
Asthma
Atopic dermatitis
Atrial fibrillation
Atrioventricular conduction
Attention deficit hyperactivity disorder
Autism
Bilirubin levels
Biochemical measures
Bipolar disorder
Blood pressure
Body mass index
Bone mineral density
Brain structure
Breast cancer
C-reactive protein
Caffeine consumption
Calcium levels
Cannabis dependence
Cardiac hypertrophy
Cardiac repolarization
Cardiovascular disease risk factors
Carotid atherosclerosis in HIV infection
Carotid intima media thickness
Celiac disease
Chemerin levels
Cholesterol, total
Chronic lymphocytic leukemia
Cleft lip
Cognitive performance
Cognitive test performance
Colorectal cancer
Common traits (Other)
Conduct disorder
Corneal structure
Coronary heart disease
Crohn's disease
D-dimer levels
Depression--quantitative trait
Diabetes related insulin traits
Diabetic retinopathy
Echocardiographic traits
Electrocardiographic conduction measures
Endometriosis
Factor VII
Fasting plasma glucose
HDL cholesterol
HDL Cholesterol - Triglycerides (HDLC-TG)
Heart failure
Height
Hematological and biochemical traits
Hemostatic factors and hematological phenotypes
Hip geometry
Hippocampal atrophy
HIV-1 control
Hoarding
Hyperactive-impulsive symptoms
Hypertension
IgE grass sensitization
Immunoglobulin A
Inattentive symptoms
Inflammatory bowel disease
Information processing speed
Iron status biomarkers
LDL cholesterol
Lipid metabolism phenotypes
Liver enzyme levels
Major depressive disorder
Menarche (age at onset)
Menarche and menopause (age at onset)
Metabolic syndrome
Metabolic traits
Metabolite levels
Morbidity-free survival
Multiple sclerosis
Myocardial infarction
Nevirapine-induced rash
Non-alcoholic fatty liver disease
Nonsyndromic cleft lip with or without cleft palate
Optic disc size (rim)
Ovarian cancer
Ovarian reserve
P-tau181p
Panic disorder
Parkinson's disease
Peripheral artery disease
Permanent tooth development
Personality dimensions
Phospholipid levels (plasma)
PR interval
Premature ovarian failure
Primary biliary cirrhosis
Primary tooth development
Prostate cancer
Psoriasis
Pulmonary function
QT interval
Quantitative traits
Reasoning
Renal cell carcinoma
Response to antidepressants
Response to antineoplastic agents
Response to antipsychotic therapy
Response to antipsychotic treatment
Response to citalopram treatment
Response to hepatitis C treatment
Response to statin therapy
Resting heart rate
Restless legs syndrome
Rheumatoid arthritis
RR interval (heart rate)
Schizophrenia
Self-rated health
Sleep duration
Smoking behavior
Sphingolipid levels
Stroke
Subclinical atherosclerosis traits (other)
Sudden cardiac arrest
Systemic lupus erythematosus
Systemic sclerosis
T-tau
Tanning
Temperament-related traits
Tonometry
Triglycerides
Triglycerides-Blood Pressure (TG-BP)
Type 1 diabetes
Type 2 diabetes
Ulcerative colitis
Urate levels
Urinary albumin excretion
Vaccine-related adverse events
Waist circumference
Waist-hip ratio
Weight
White blood cell types
Working memory

339 intergenic GWAS linkage blocks, covering a total of ~50 Mbp of the genome (~1.7%).

High global enrichment for transcription from captured GWAS regions

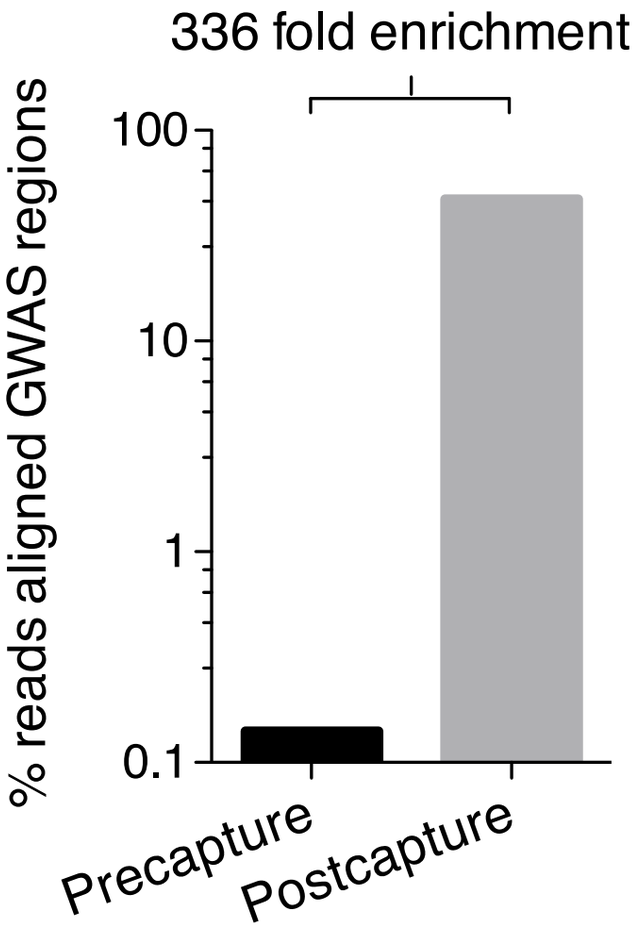


High global enrichment for transcription from captured GWAS regions



Comparing the percentage of mapped reads from targeted regions pre- and post-capture

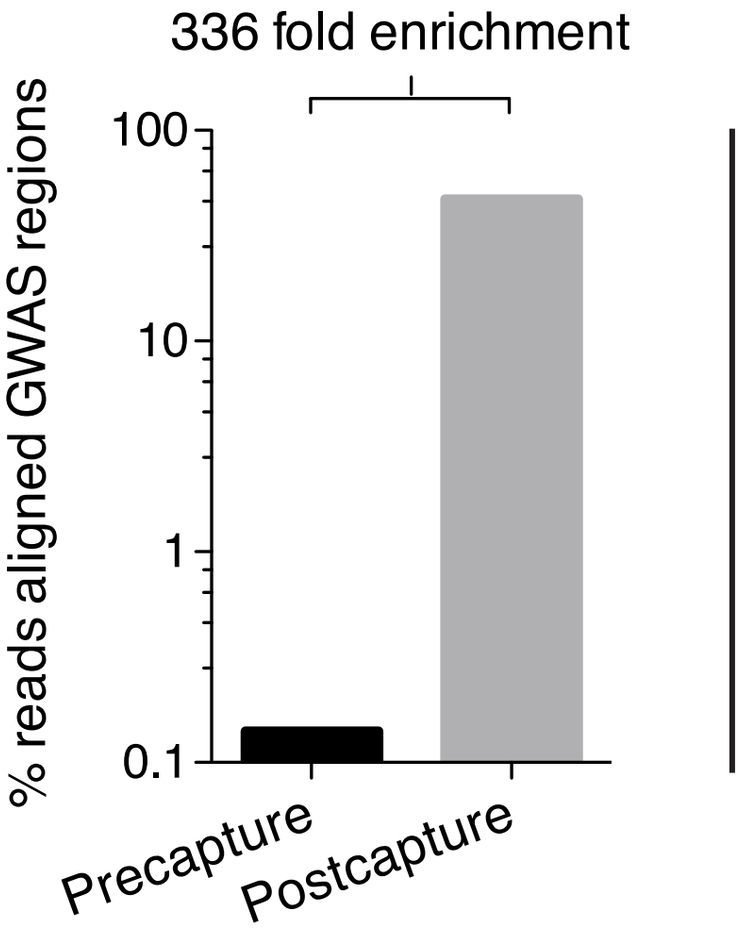
High global enrichment for transcription from captured GWAS regions



Comparing the percentage of mapped reads from targeted regions pre- and post-capture

~0.2% of the RNA in pre-capture library now comprises ~60% in the post-capture library.

High global enrichment for transcription from captured GWAS regions

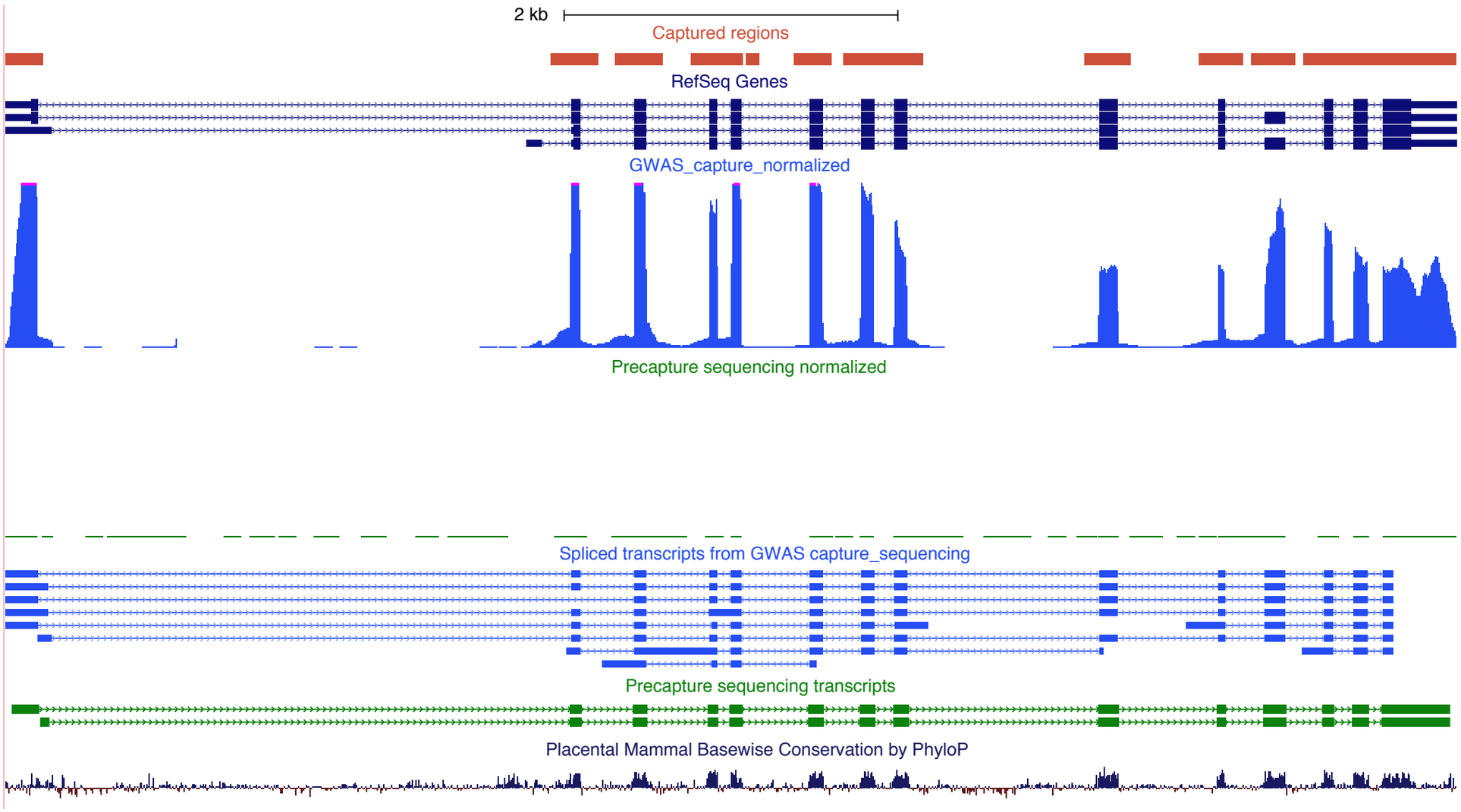


Comparing the percentage of mapped reads from targeted regions pre- and post-capture

~0.2% of the RNA in pre-capture library now comprises ~60% in the post-capture library.

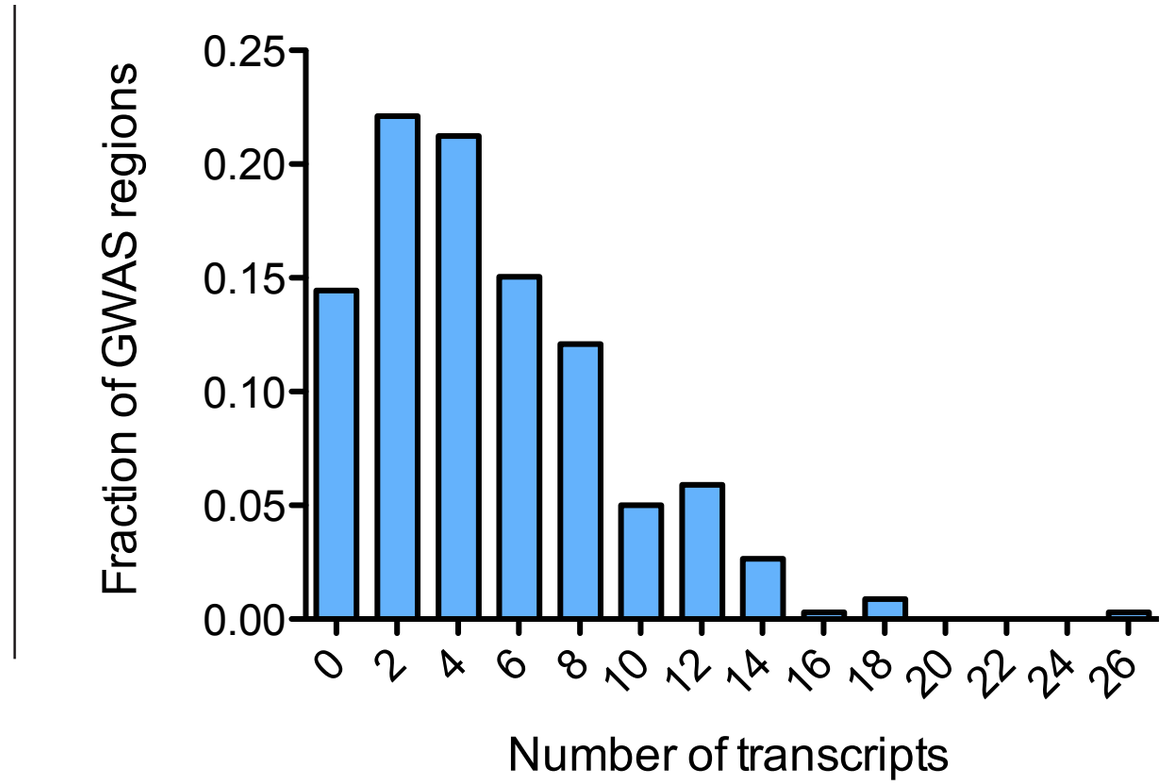
336 fold enrichment from CaptureSeq

Sequencing based enrichment: HMBS control loci



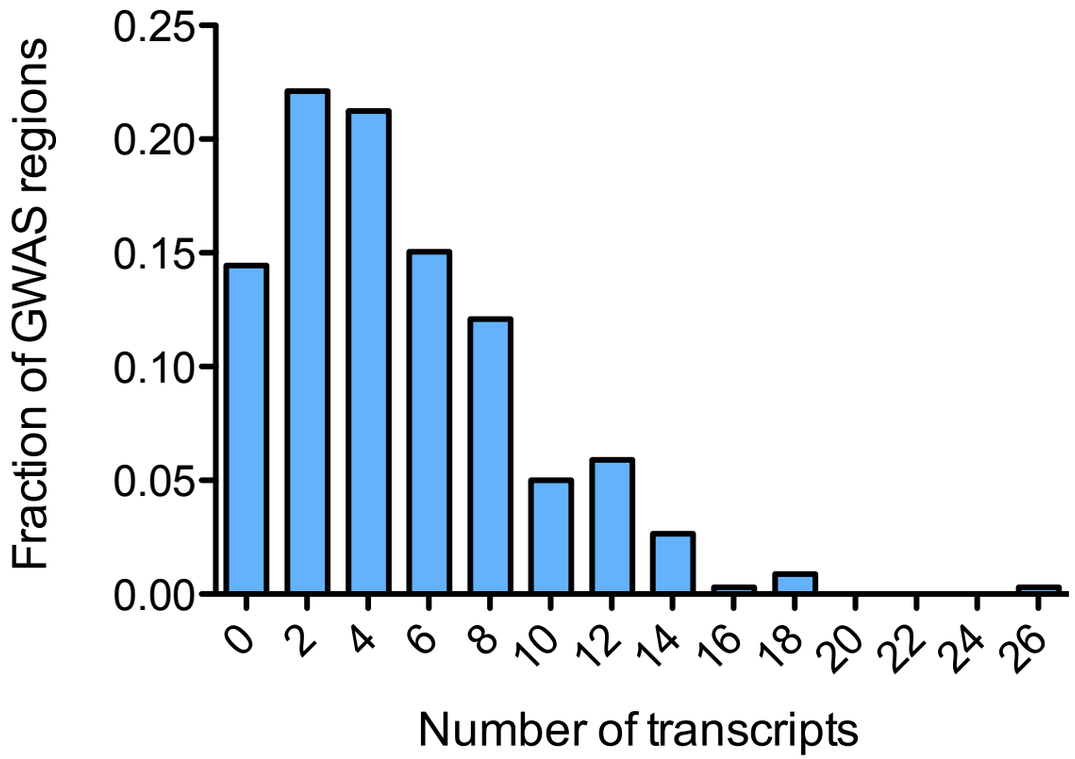
- HMBS is a control loci that we capture
- Sequencing confirms large enrichment measured globally and by qPCR
- Assemble realistic transcripts from Capture sequencing

Extensive transcription within intergenic GWAS regions



Extensive transcription within intergenic GWAS regions

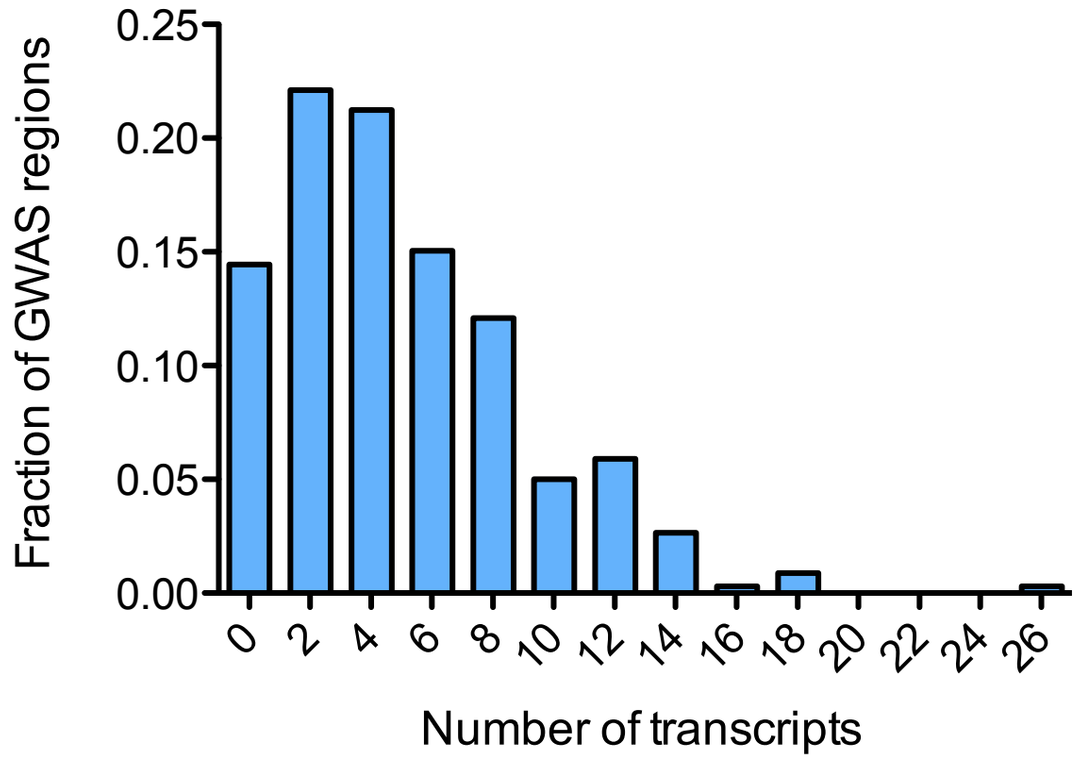
85% of 339 captured GWAS regions contained spliced transcripts.



Extensive transcription within intergenic GWAS regions

85% of 339 captured GWAS regions contained spliced transcripts.

Median of 4 spliced transcripts per locus

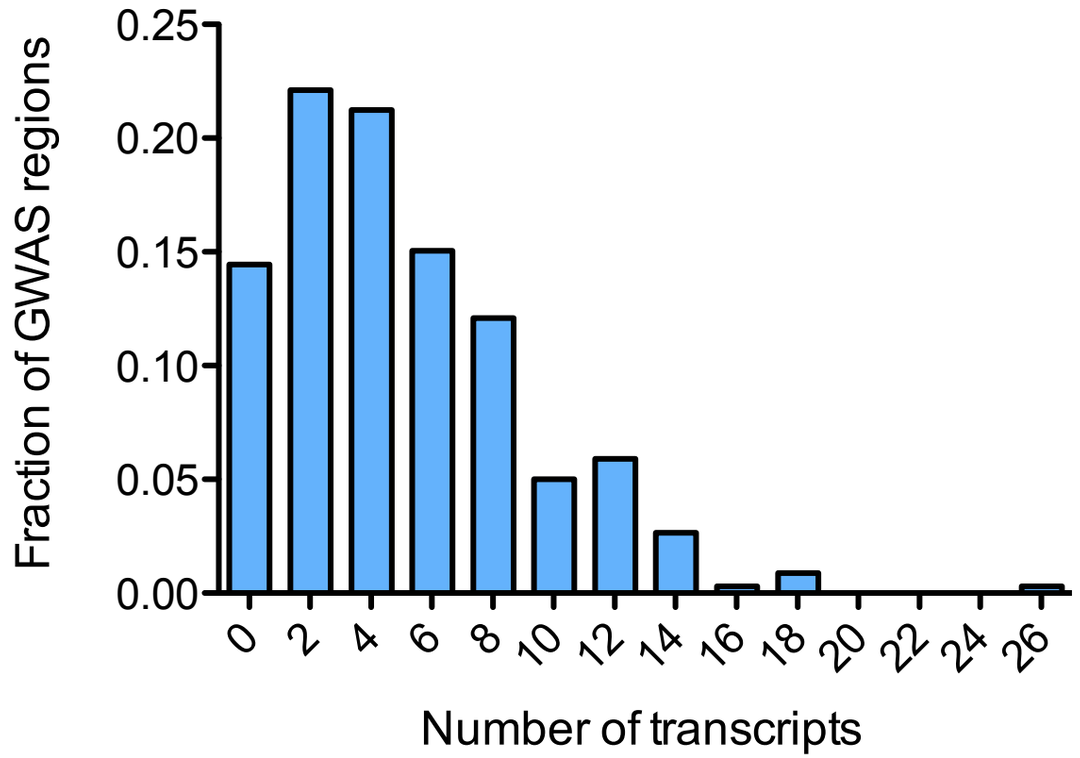


Extensive transcription within intergenic GWAS regions

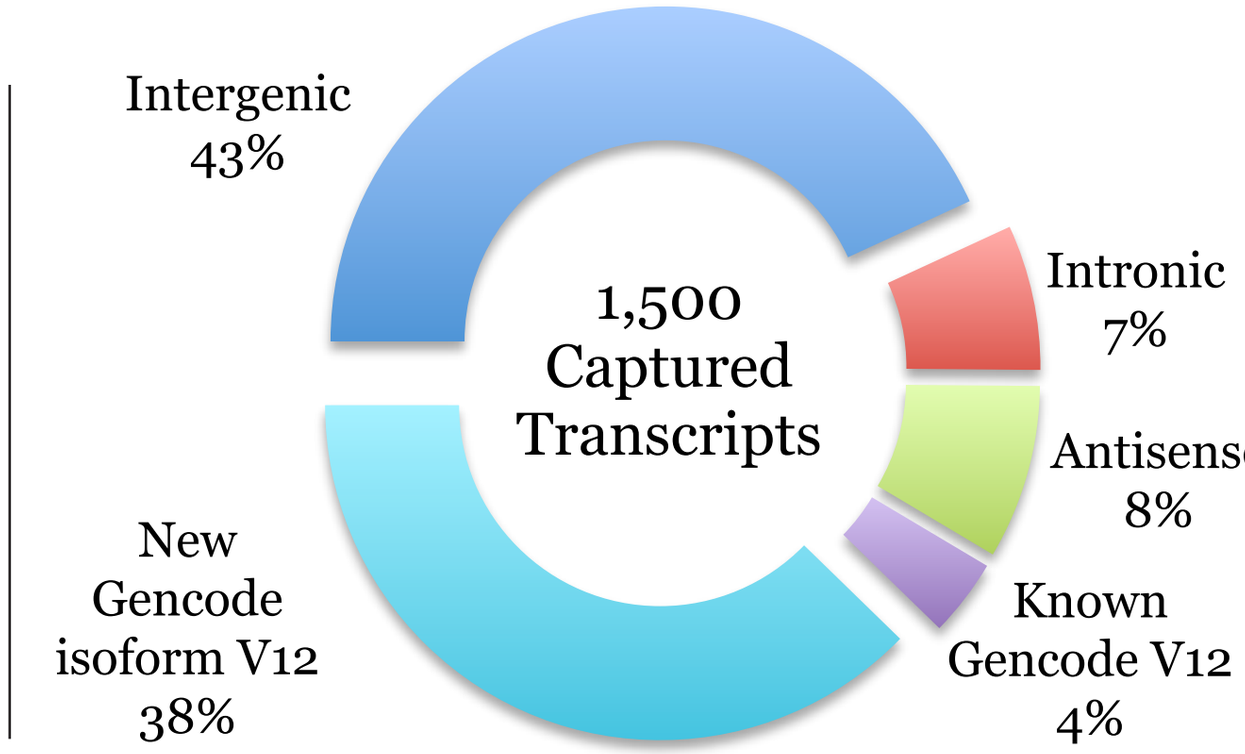
85% of 339 captured GWAS regions contained spliced transcripts.

Median of 4 spliced transcripts per locus

Spliced transcripts covered a total of 55.6 Mb with many transcripts transcribed into, or out of, the captured regions.

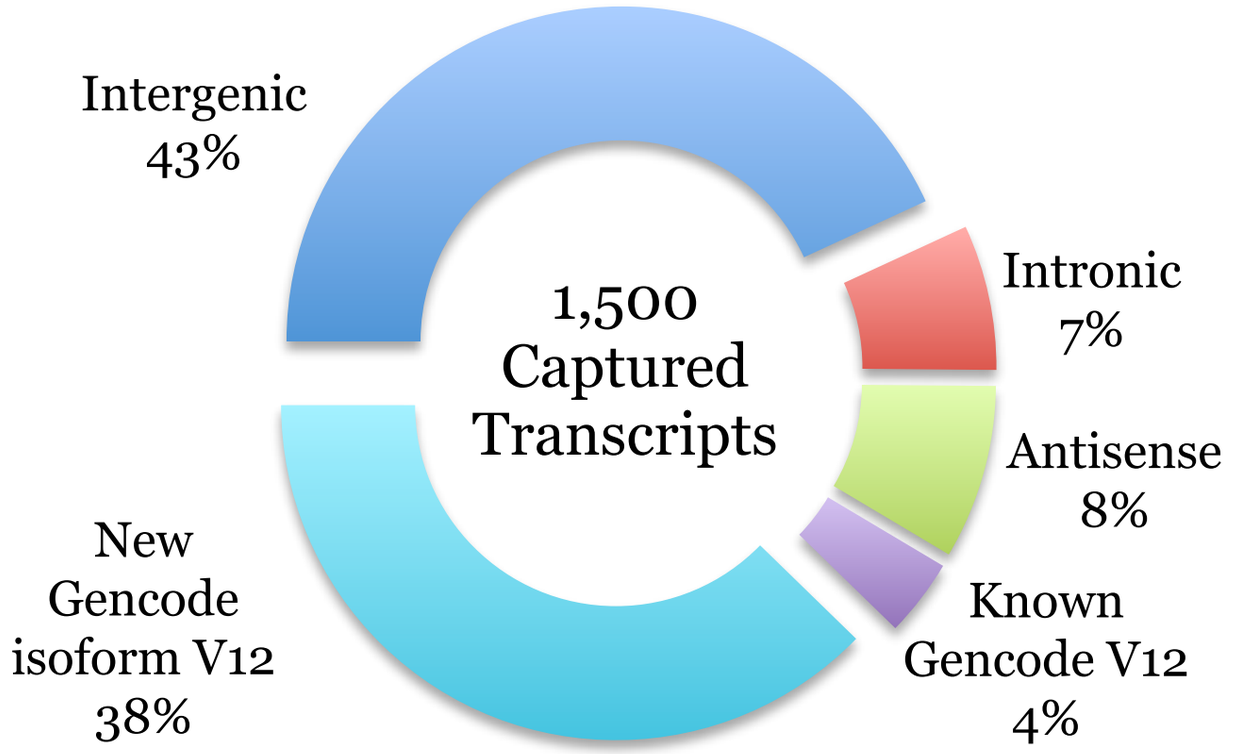


1500 transcripts identified in “intergenic” GWAS regions



1500 transcripts identified in “intergenic” GWAS regions

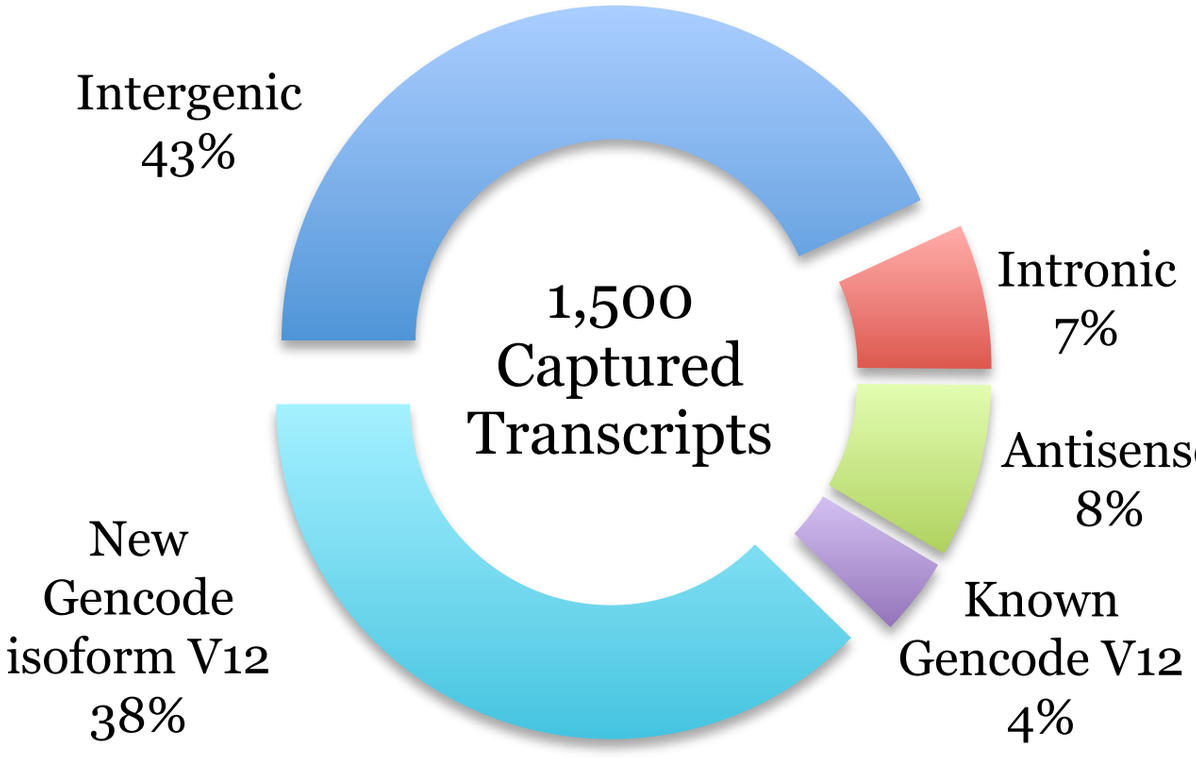
1,500 transcripts were identified (compared to 110 transcript in pre-capture).



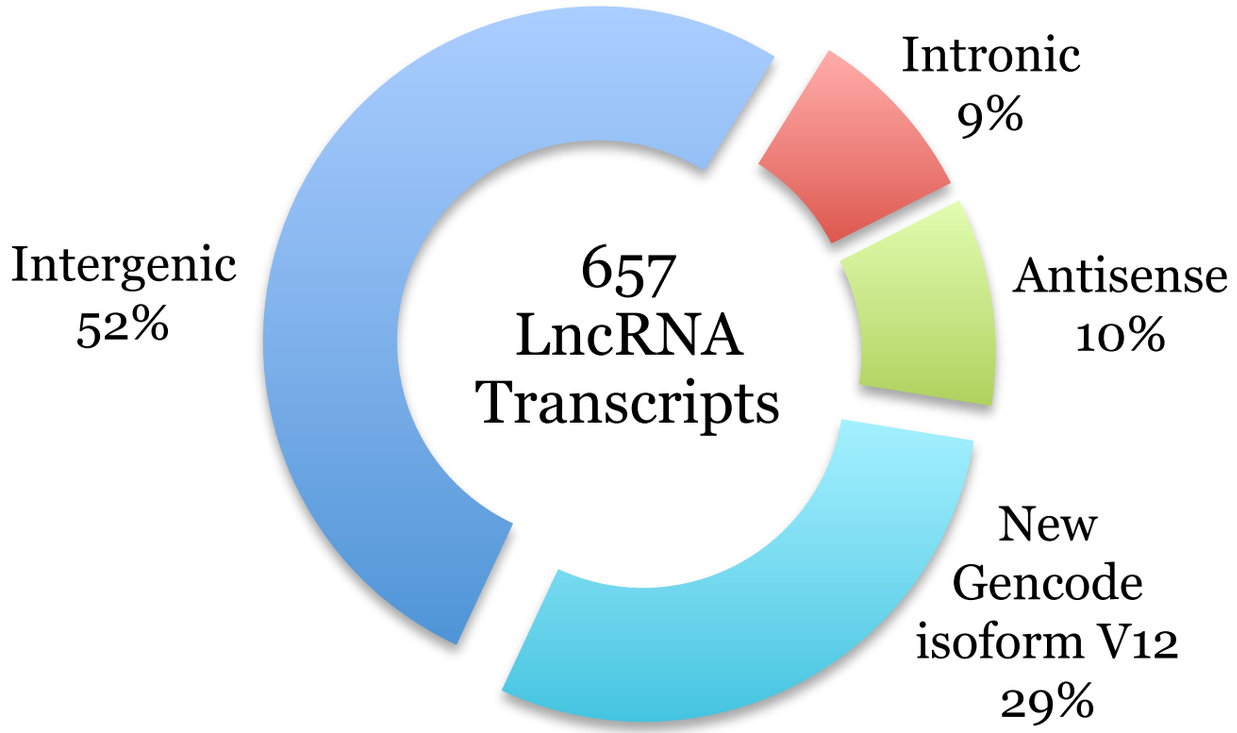
1500 transcripts identified in “intergenic” GWAS regions

1,500 transcripts were identified (compared to 110 transcript in pre-capture).

Majority of captured transcripts are entirely novel or contain novel exons of distal genes.

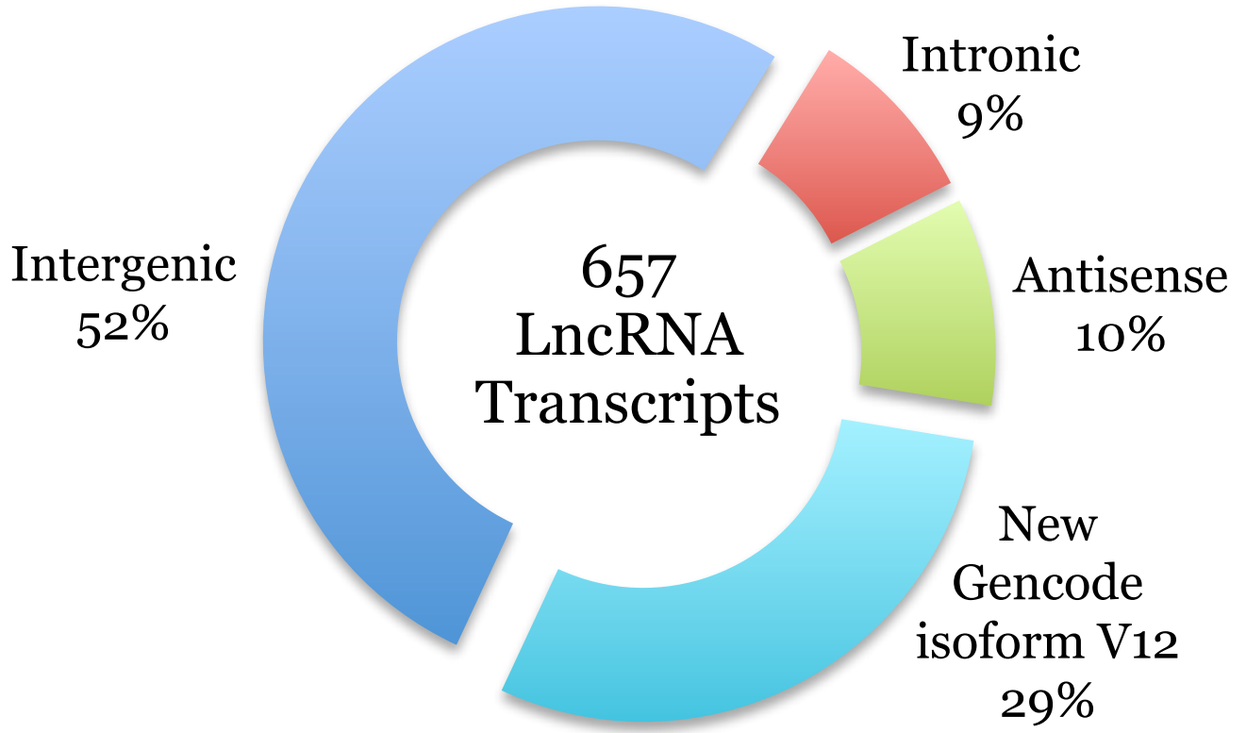


650 novel lncRNAs identified in “intergenic” GWAS regions



650 novel lncRNAs identified in “intergenic” GWAS regions

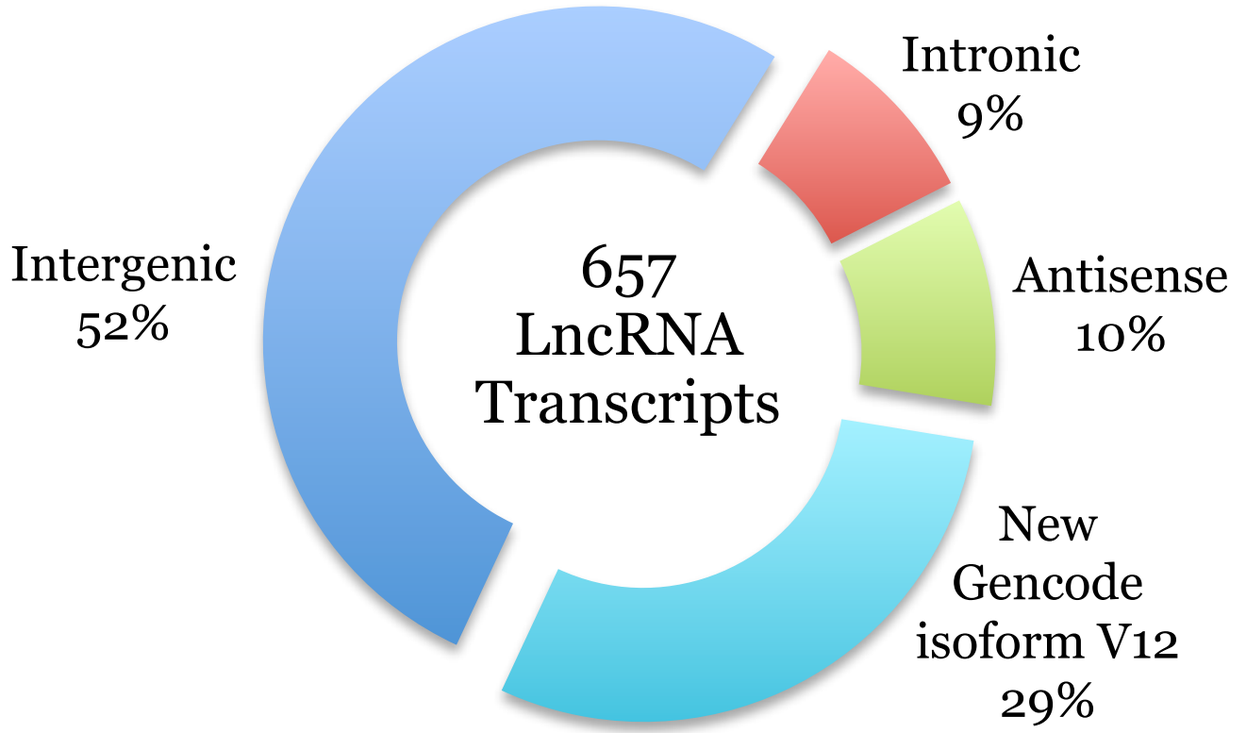
Stringent filtering identifies 657 new lncRNAs in 369 loci



650 novel lncRNAs identified in “intergenic” GWAS regions

Stringent filtering identifies 657 new lncRNAs in 369 loci

Vast majority are completely novel

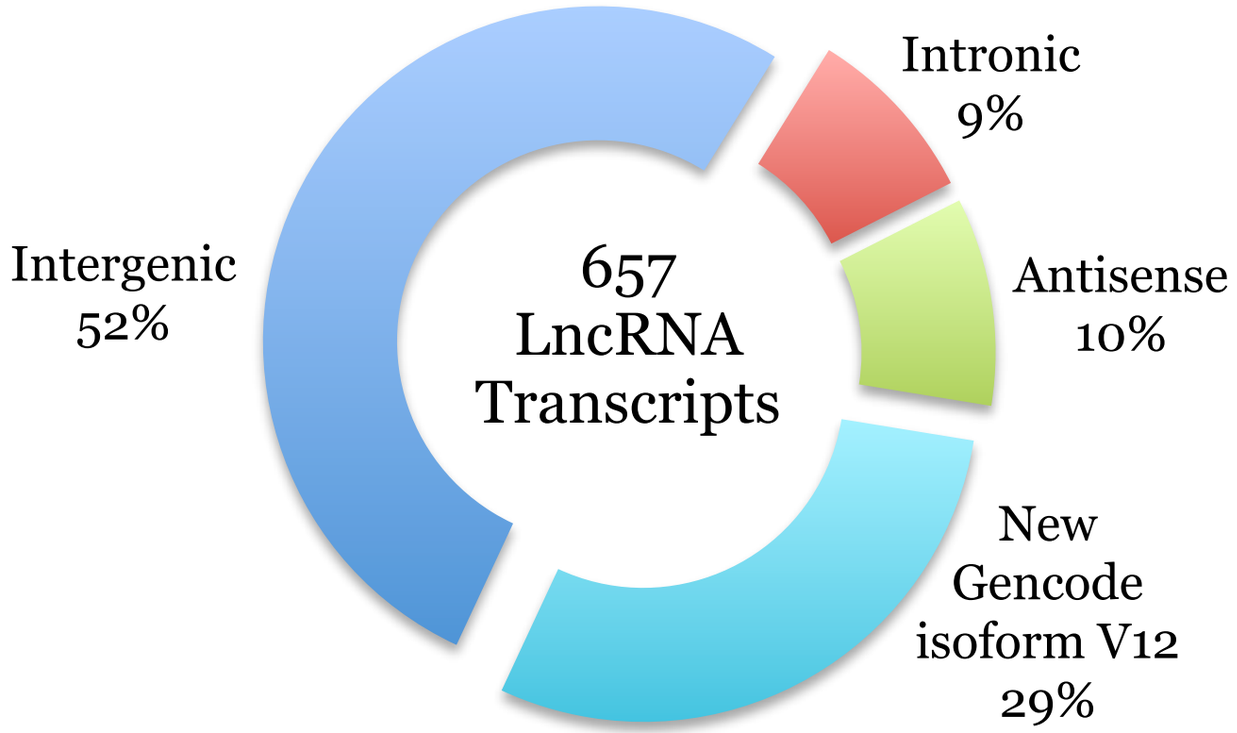


650 novel lncRNAs identified in “intergenic” GWAS regions

Stringent filtering identifies 657 new lncRNAs in 369 loci

Vast majority are completely novel

Remainder of ~1500 transcripts are novel coding gene isoforms and transcripts of uncertain coding potential.

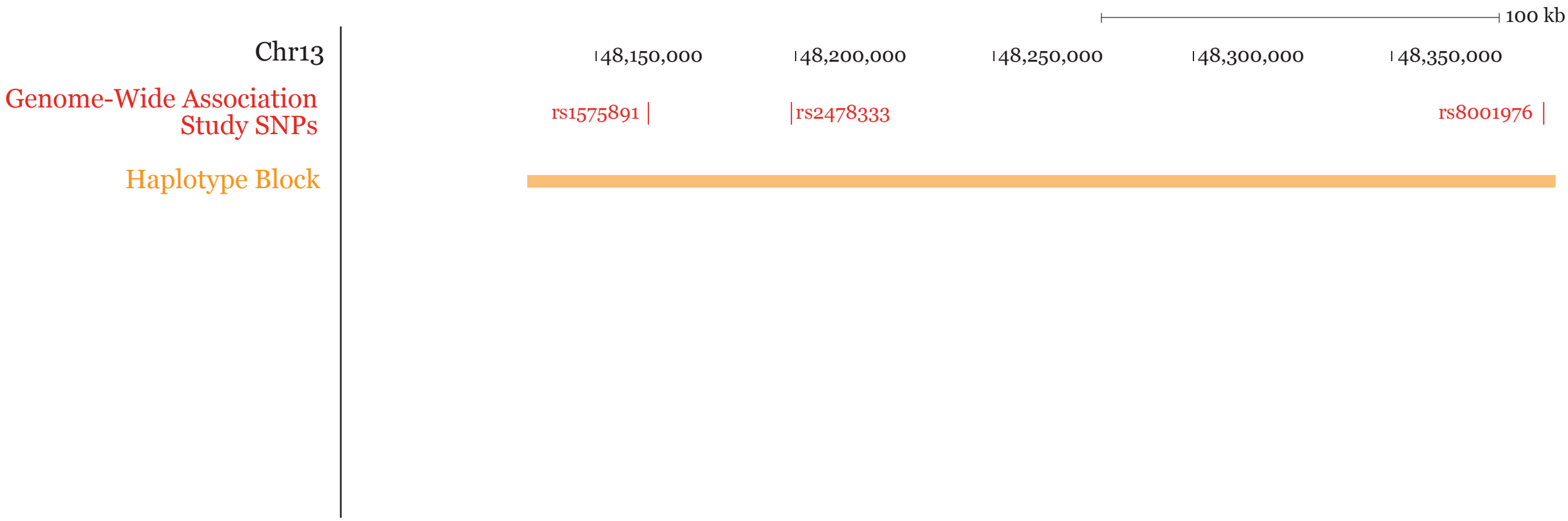


Identifying novel human genes with capture seq

Human loci associated with heart QT length, but doesn’t contain any known genes.

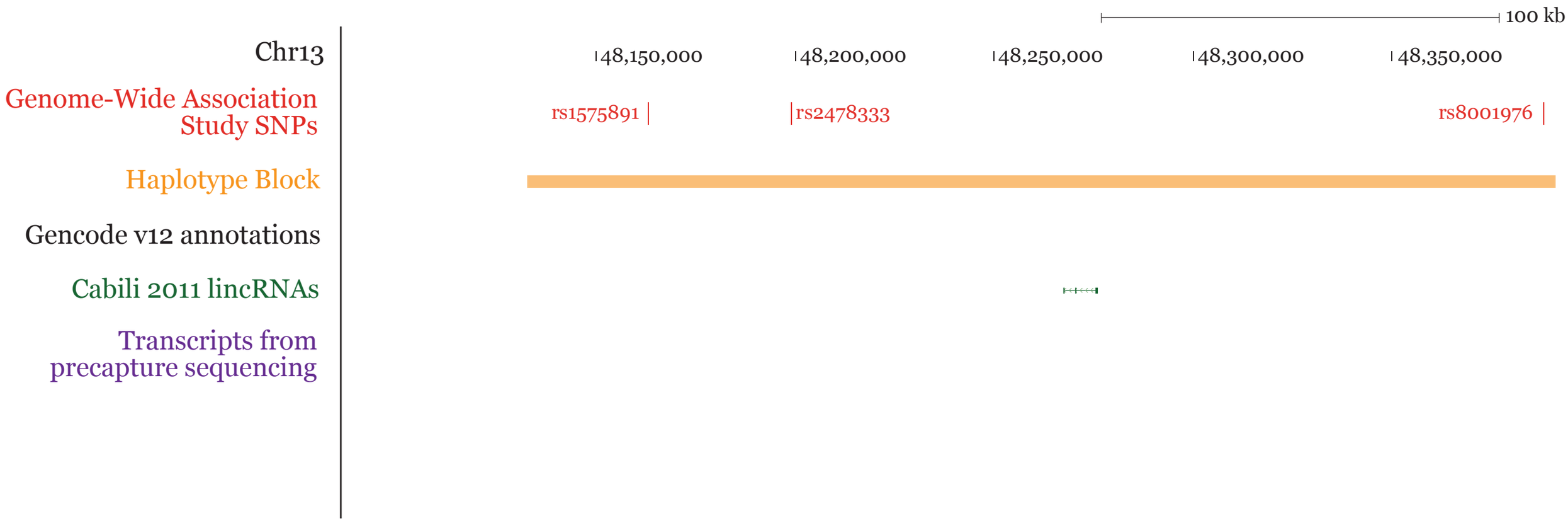
Identifying novel human genes with capture seq

Human loci associated with heart QT length, but doesn't contain any known genes.



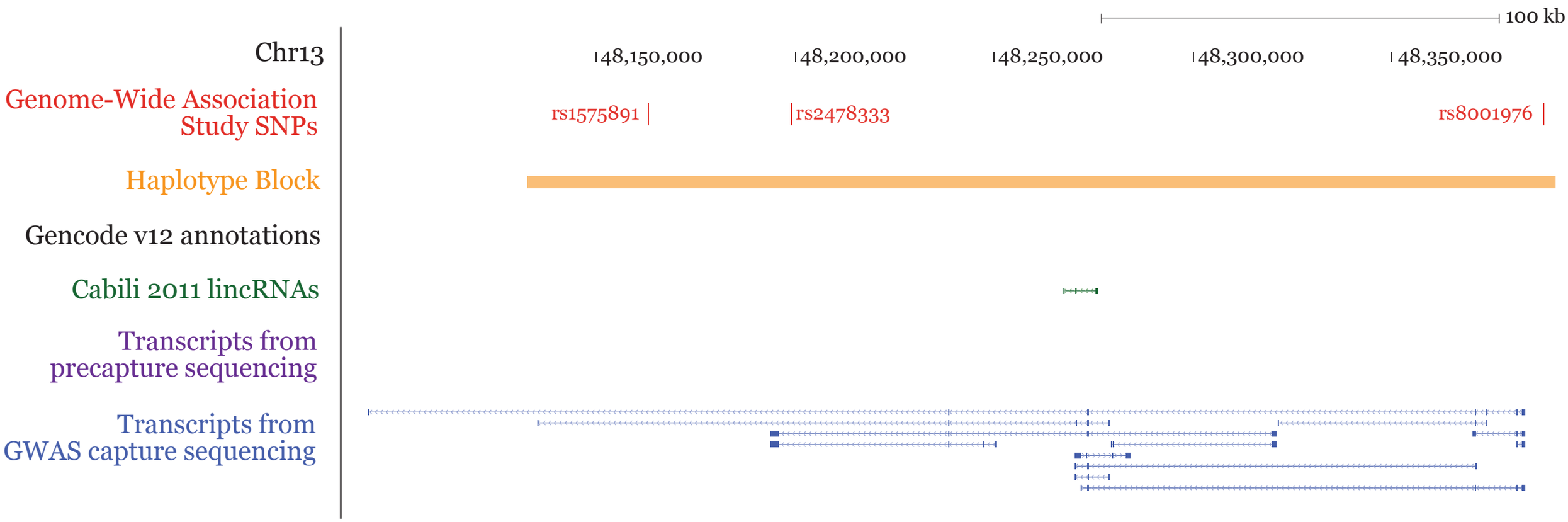
Identifying novel human genes with capture seq

Human loci associated with heart QT length, but doesn't contain any known genes.



Identifying novel human genes with capture seq

Human loci associated with heart QT length, but doesn't contain any known genes.



Capture sequencing finds many new transcripts, these can now be functionally tested.

Identifying novel human genes with capture seq

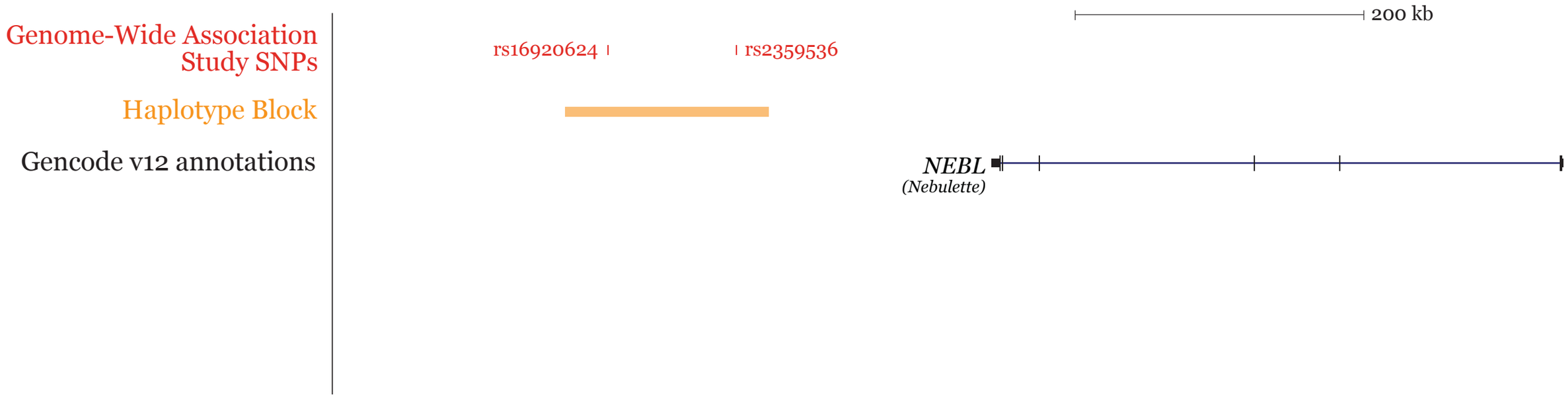
GWAS regions contains novel 5’ or 3’ extensions of known genes.

Identifying novel human genes with capture seq

GWAS regions contains novel 5’ or 3’ extensions of known genes.

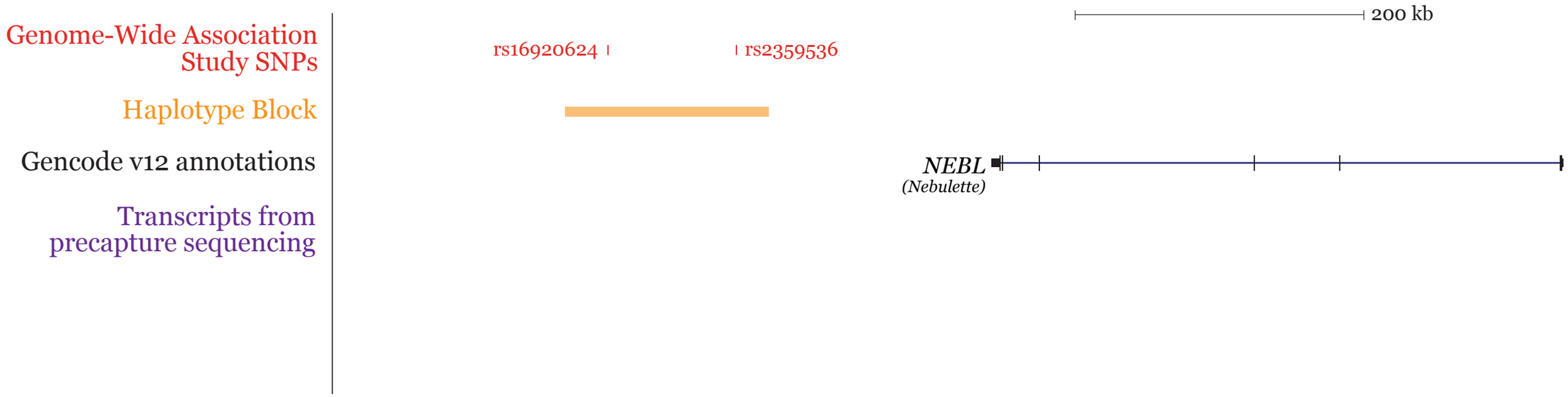
Identifying novel human genes with capture seq

GWAS regions contains novel 5’ or 3’ extensions of known genes.



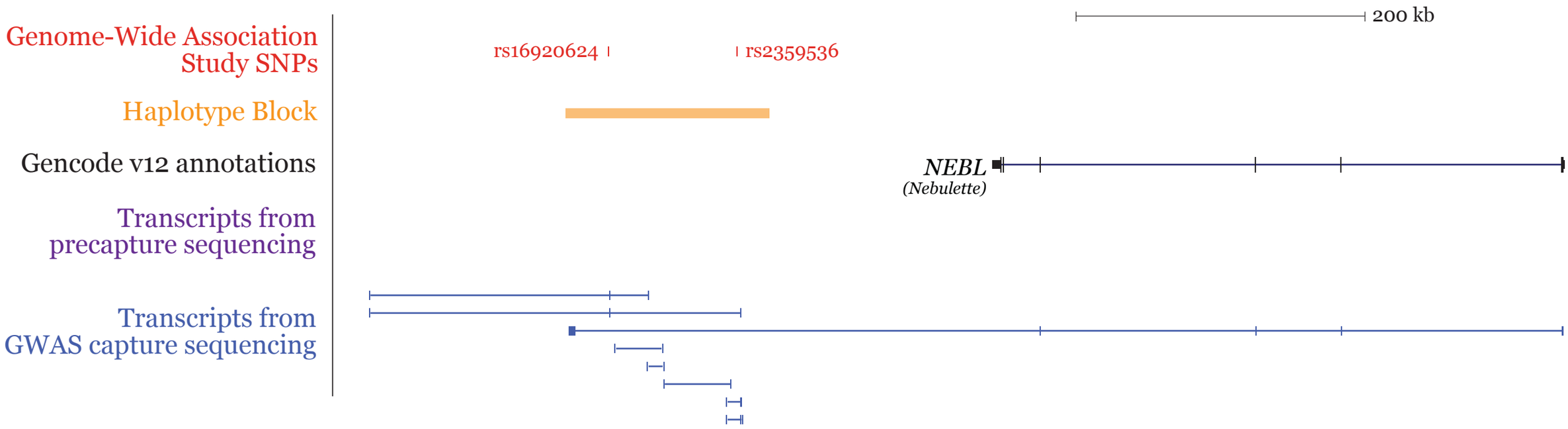
Identifying novel human genes with capture seq

GWAS regions contains novel 5’ or 3’ extensions of known genes.



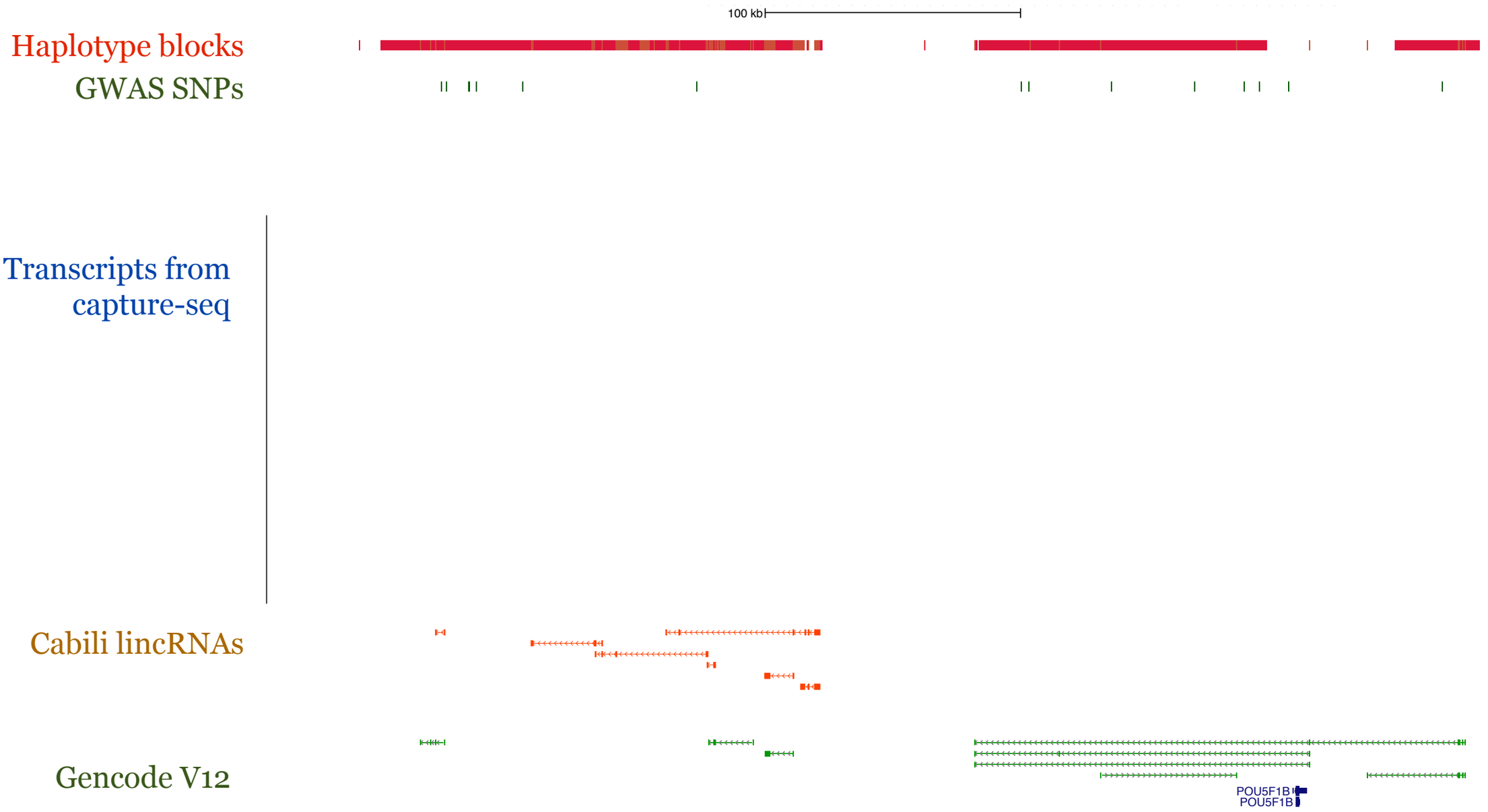
Identifying novel human genes with capture seq

GWAS regions contains novel 5’ or 3’ extensions of known genes.



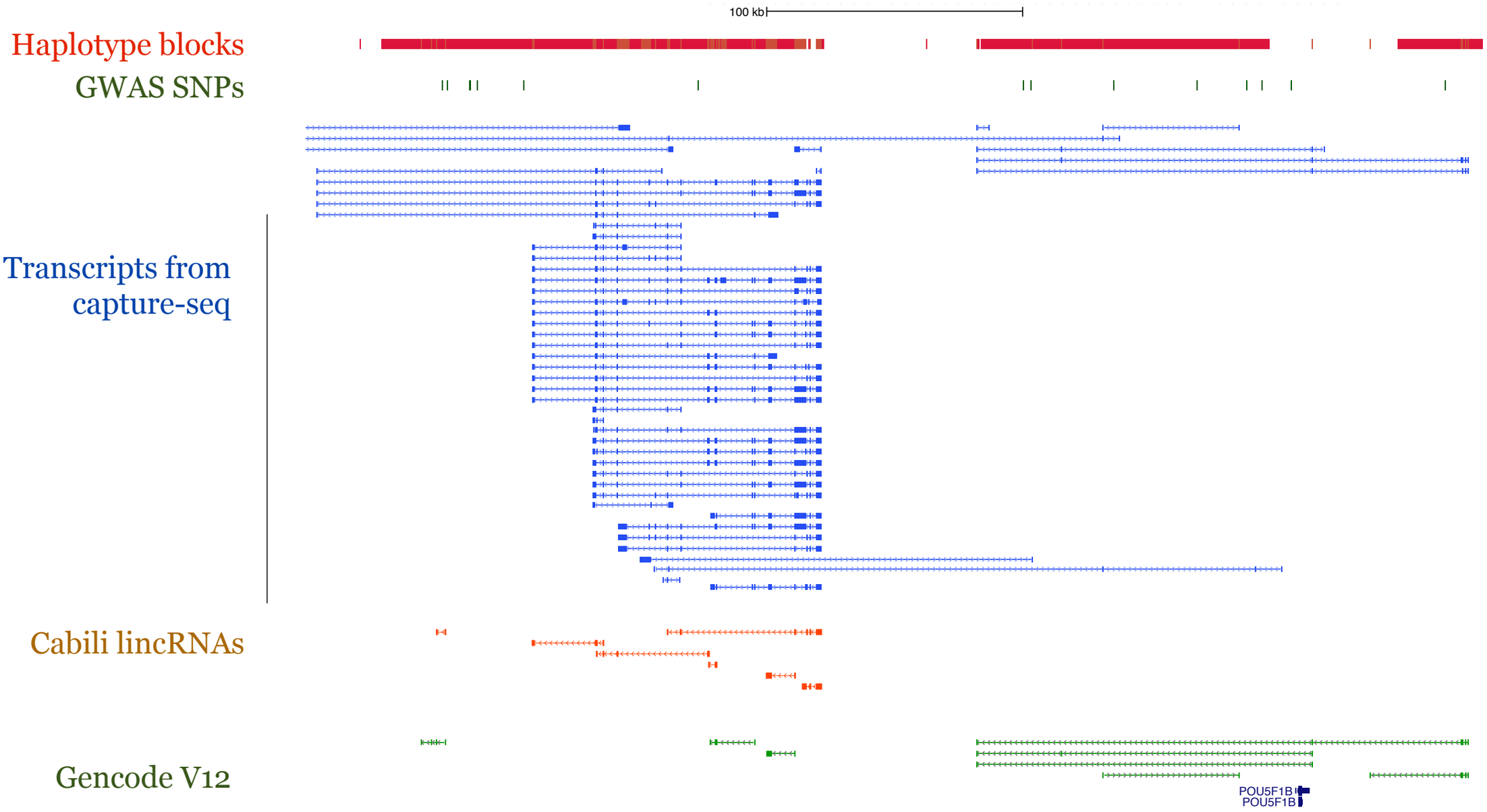
- Captured GWAS region contains a novel isoform of the coding gene Nebulette (NEBL).
- A very recently annotated putative miRNA is present in the GWAS region on +ve strand. Likely we have also identified its host transcript(s).

Capture seq resolves fragmented annotations



- Region of chr8 associated with prostate cancer by a number of studies.

Capture seq resolves fragmented annotations



- Region of chr8 associated with prostate cancer by a number of studies.
- Capture sequencing combines previously annotated lncRNA fragments in complex alternatively spliced loci

Capture-seq vs Hangauer 2013 lincRNA catalogue

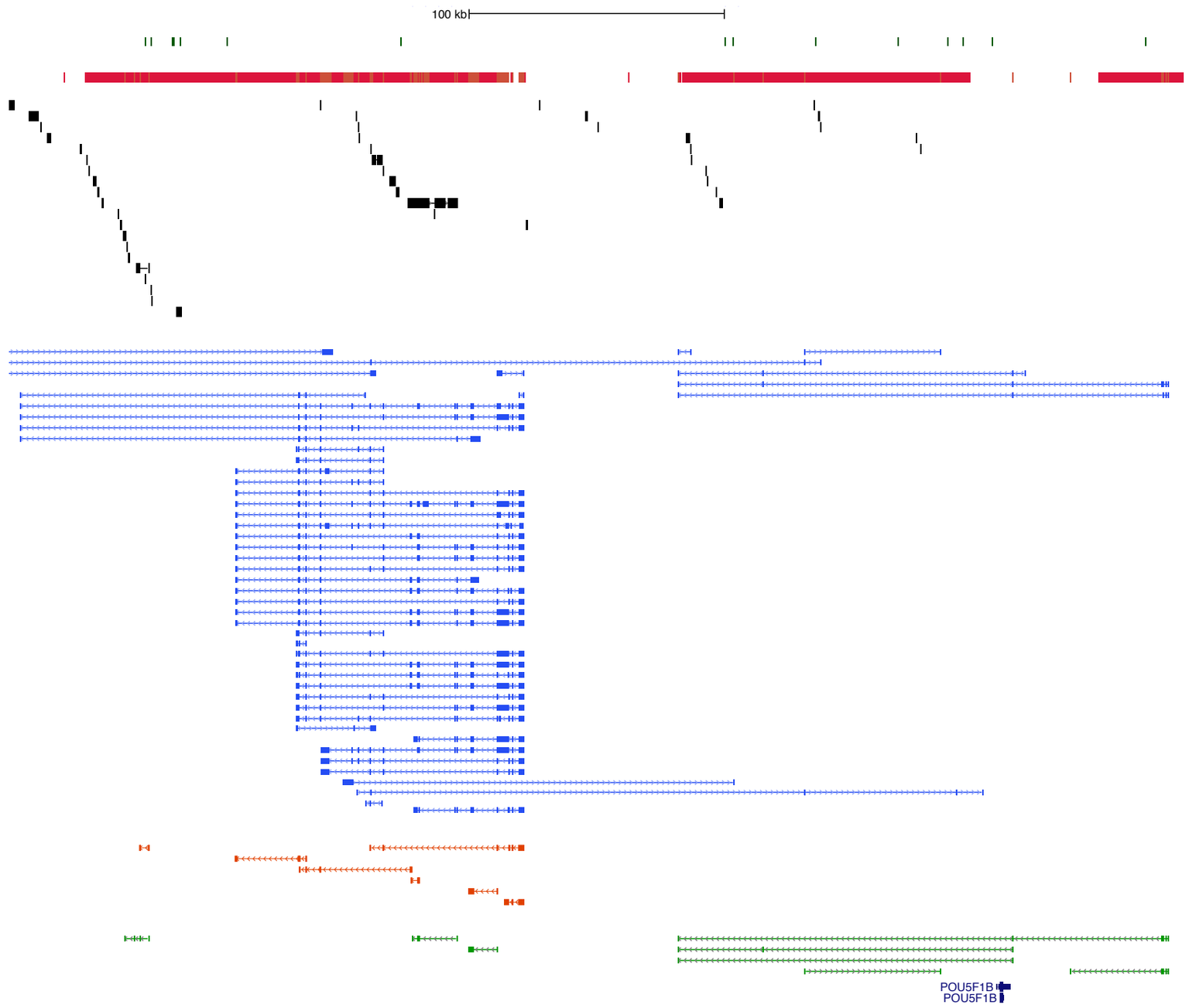
GWAS SNPs
Haplotype blocks

Hangauer 2013
lincRNA catalogue

Transcripts from
capture-seq

Cabili lincRNAs

Gencode V12

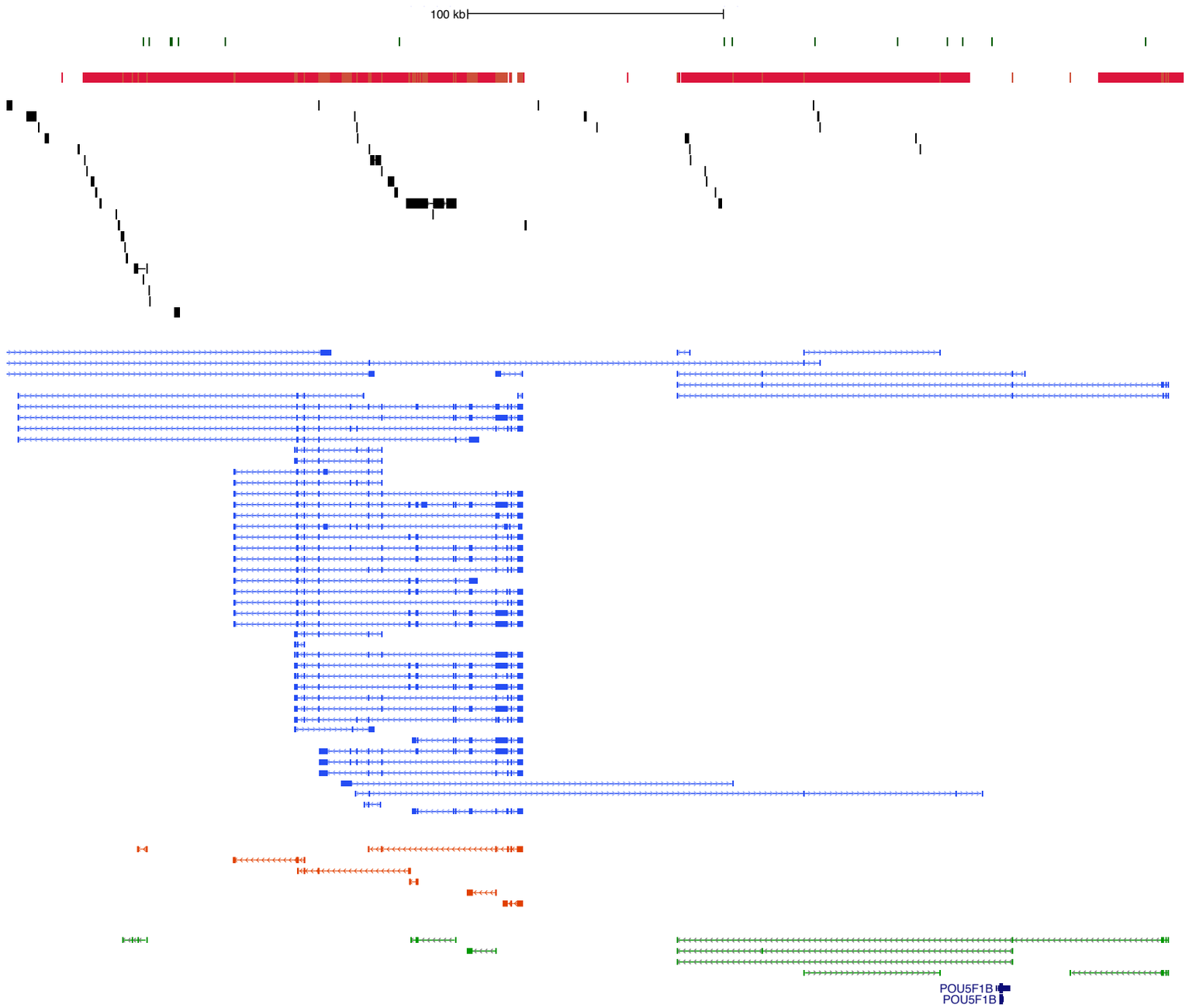


Capture-seq vs Hangauer 2013 lincRNA catalogue

GWAS SNPs
Haplotype blocks

Hangauer 2013
lincRNA catalogue

Transcripts from
capture-seq



Cabili lincRNAs

Gencode V12

Capture-seq outperforms other novel gene catalogues

Summary: Intergenic GWAS capture sequencing

Summary: Intergenic GWAS capture sequencing

- Capture-seq allows unprecedented depth of coverage and enriches transcripts standard RNA-seq cannot identify.

Summary: Intergenic GWAS capture sequencing

- Capture-seq allows unprecedented depth of coverage and enriches transcripts standard RNA-seq cannot identify.
- We find extensive transcription within and across “intergenic” GWAS regions, identifying ~1500 mostly novel transcripts.

Summary: Intergenic GWAS capture sequencing

- Capture-seq allows unprecedented depth of coverage and enriches transcripts standard RNA-seq cannot identify.

- We find extensive transcription within and across “intergenic” GWAS regions, identifying ~1500 mostly novel transcripts.

- These novel transcripts provide candidates to explain the traits associated with some of these regions.

Summary: Intergenic GWAS capture sequencing

- Capture-seq allows unprecedented depth of coverage and enriches transcripts standard RNA-seq cannot identify.

- We find extensive transcription within and across “intergenic” GWAS regions, identifying ~1500 mostly novel transcripts.

- These novel transcripts provide candidates to explain the traits associated with some of these regions.

- Investigating intergenic GWAS regions with capture seq is a good method for identifying potentially functional lncRNAs.

Acknowledgements

Garvan Institute
John Mattick
Marcel Dinger
Tim Mercer

Institute for Molecular Bioscience
Ryan Taft
Jo Crawford
Kelin Ru

Roche
Jeffrey Jeddeloh

Mater Medical Research Institute
Daniel Gerhardt

Acknowledgements

Garvan Institute
John Mattick
Marcel Dinger
Tim Mercer

Institute for Molecular Bioscience
Ryan Taft
Jo Crawford
Kelin Ru

Roche
Jeffrey Jeddeloh

Mater Medical Research Institute
Daniel Gerhardt