



Farmed animal genomes – status and where next

Alan L. Archibald, David W. Burt

The Roslin Institute, University of Edinburgh, UK

Brian P. Dalrymple

CSIRO, Australia





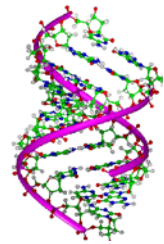
Acknowledgements

- Swine Genome Sequencing Consortium
- International Sheep Genome Consortium
- Chicken and other avian genome consortia
- Many others
- Funders
 - Many sources, including
 - EC FP7 Quantomics-222664
 - EC FP7 3SR-245140



From Sequence to Consequence

Tools for the Exploitation of Livestock Genomes



1953
Watson and
Crick



1977
DNA
sequenced
! X174
5,386 nt

1990
Human
Genome
Project
launched



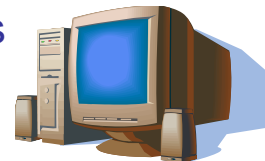
1991
PiGMAP
project
starts
'Halothane'
gene test

2001
Draft human
genome sequence



1920s and 30s
Fisher, Lush
and others
Population
Genetics

1970s +
Advances in
quantitative
analysis



1990s +
Quantitative
trait locus
(QTL)
mapping

2001
Genomic selection
proposed



Reference genome sequence as a key resource and framework for biological research

- Genetics
 - Variation (SNPs, indels, CNVs)
 - SNP chips, Genotype-by-Sequence
 - Genome-Wide Association Studies (GWAS)
 - Genetic improvement
- Functional genomics
 - incl. physiology, immunology,.....
 - Genome-wide analysis of responses to perturbation
 - Gene expression, methylation,
 - Microarrays, Assay-by-sequence



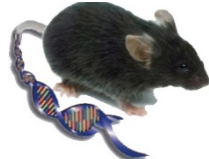
2010
Turkey genome sequenced



2009
Cattle genome sequenced



Horse genome Sequenced



Mouse genome "finished"

2008
Human 1000 Genomes Project launched



2007
Cat genome sequenced

2004
Chicken genome sequenced

2005
Dog genome sequenced

2003
Human genome sequence "finished" \$3 billion



2002
Mouse draft genome sequence

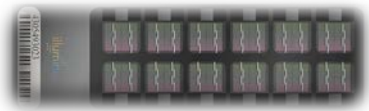


2008
Bovine 50K SNP chip

2009
Pig 60K SNP chip

Sheep 60K SNP chip

2010
750K bovine SNP chips

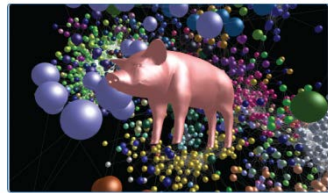




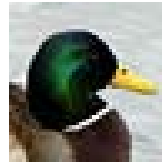
2013
Goat genome
sequenced

2013 onwards
Animal
ENCODE

BMC Biology
University Journal of Biology



2012
Pig gene
expression atlas



2013
Duck genome
sequenced



2013
Sheep genome
sequenced

2013 onwards
Genotype-by-
sequence



2012
Pig genome
sequenced



2012
Chicken 600K
SNP chip



2013
Salmon
SNP chip

Fish: Tilapia, Cod, Salmon,.....





Genome variation

- From a reference to multiple genomes per species
- Enables
 - Discovery of SNP and structural variation (& SNP chips)
 - Analysis of natural and artificial selection
 - Identification of causal variation
- Visualising genetic variation (Ensembl – Variation)
- Predicting consequences (e.g. Variant Effect Predictor, SIFT,...)

Multiple genomes

- Human 1000 Genomes Project
 - ~4-6x coverage / individual
 - revealing genetic burden
 - ~1-200 potential Loss of Function mutations per person
- Human genetics studies
 - 10's of thousands per study
 - ICQG 2012
 - 30K sequenced genomes in a study

Multiple animal genomes

- Pooled samples
 - 10-15x coverage
 - Chickens, cattle, pigs
 - SNP discovery
 - Signatures of selection
 - Signatures of domestication
- Individual genomes
 - 4-10x coverage
 - €3,000 per genome

Multiple animal genomes


- 1000 Bull Genomes Project
 - Collaborative, Cloud data repository
 - Nnn bulls, average coverage ~11x
 - Data analysis cycles for genomic prediction
- Pigs
 - Groenen (Wageningen) ~300 individual pigs
 - Korean ~60 individual pigs
 - China ?? Pigs
- Sheep
 - ISGC 75 individual sheep
- Chickens
 - 10's of individuals (e.g. 10 individual J line brown egg layers)

Visualising genome variation

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

 [More about variation in Ensembl](#)

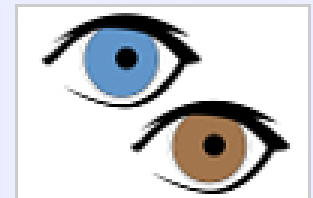
 [Download all variants \(GVF\)](#)

 [Variant Effect Predictor](#)



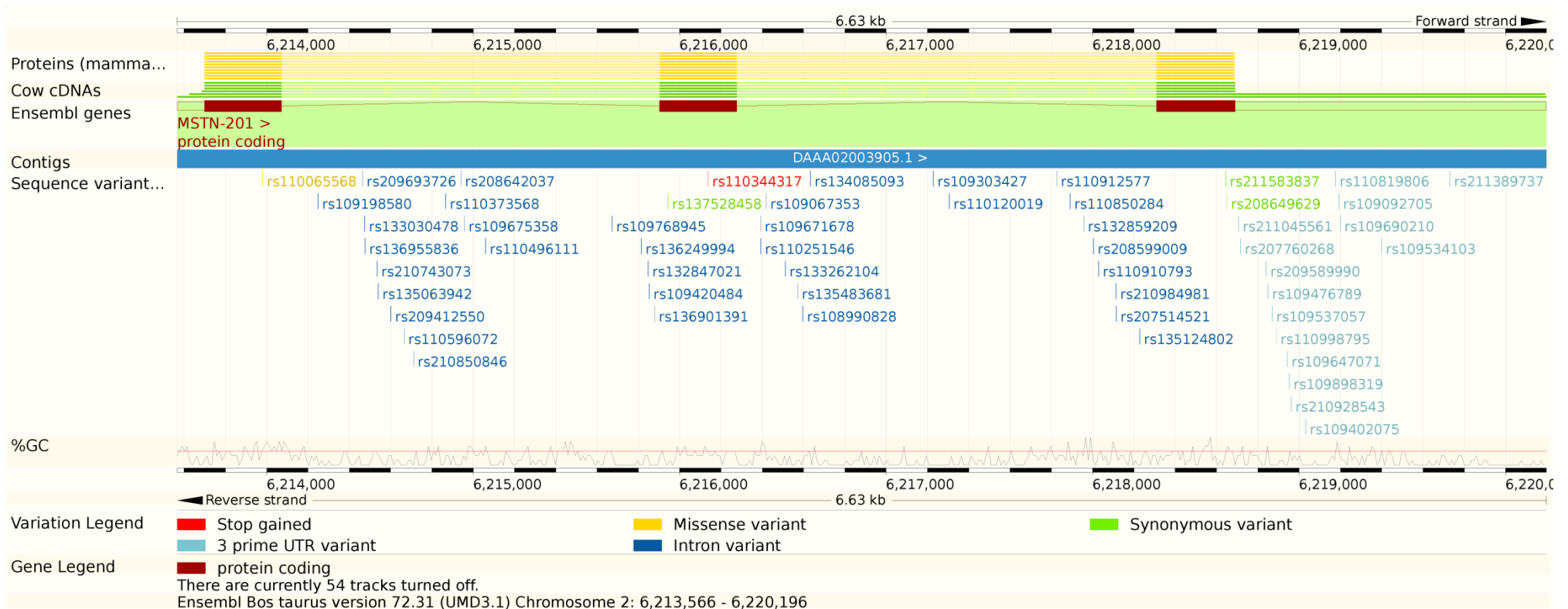
```
ATCGAGCT  
ATCCAGCT  
ATCGAGAT
```

Example variant



Example phenotype

GDF8 - SNPs



DGAT1 - variation

Summary of variation consequences in ENSBTAG00000026356




[Switch to tree view](#)

Show All entries		Filter	
Number of variant consequences	Type	Description	
0	- ■ Transcript ablation	A feature ablation whereby the deleted region includes a transcript feature (SO:0001893)	
0	- ■ Splice donor variant	A splice variant that changes the 2 base region at the 5' end of an intron (SO:0001575)	
0	- ■ Splice acceptor variant	A splice variant that changes the 2 base region at the 3' end of an intron (SO:0001574)	
0	- ■ Stop gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript (SO:0001587)	
0	- ■ Frameshift variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three (SO:0001589)	
0	- ■ Stop lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript (SO:0001578)	
0	- ■ Initiator codon variant	A codon variant that changes at least one base of the first codon of a transcript (SO:0001582)	
0	- ■ Transcript amplification	A feature amplification of a region containing a transcript (SO:0001889)	
0	- ■ Inframe insertion	An inframe non synonymous variant that inserts bases into in the coding sequence (SO:0001821)	
0	- ■ Inframe deletion	An inframe non synonymous variant that deletes bases from the coding sequence (SO:0001822)	
5	Show ■ Missense variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved (SO:0001583)	
1	Show ■ Splice region variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron (SO:0001630)	
0	- ■ Incomplete terminal codon variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed (SO:0001626)	
1	Show ■ Synonymous variant	A sequence variant where there is no resulting change to the encoded amino acid (SO:0001819)	
0	- ■ Stop retained variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains (SO:0001587)	
0	- ■ Coding sequence variant	A sequence variant that changes the coding sequence (SO:0001580)	
0	- ■ Mature miRNA variant	A transcript variant located with the sequence of the mature miRNA (SO:0001620)	
0	- ■ 5 prime UTR variant	A UTR variant of the 5' UTR (SO:0001623)	

DGAT1 – missense variants

Missense variant consequences 

[\[back to top\]](#)

Show/hide columns Filter 											
ID	Chr: bp	Alleles	Class	Source	Evidence	Type	AA	AA co-ord	SIFT	Transcript	
rs134083952	14:1803973	T/C	SNP	dbSNP	-	Missense variant	F/L	370	0	ENSBTAT0000037423	
rs135329220	14:1804495	T/G	SNP	dbSNP	-	Missense variant	V/G	464	0.01	ENSBTAT0000037423	
rs137745035	14:1801941	A/G	SNP	dbSNP	-	Missense variant	T/A	187	0.04	ENSBTAT0000037423	
rs109234250	14:1802265	G/A	SNP	dbSNP		Missense variant	A/T	232	0.58	ENSBTAT0000037423	
rs109326954	14:1802266	C/A	SNP	dbSNP		Missense variant	A/E	232	1	ENSBTAT0000037423	

cf. K232E Grisart et al 2003 PNAS 101: 2398

flawed link



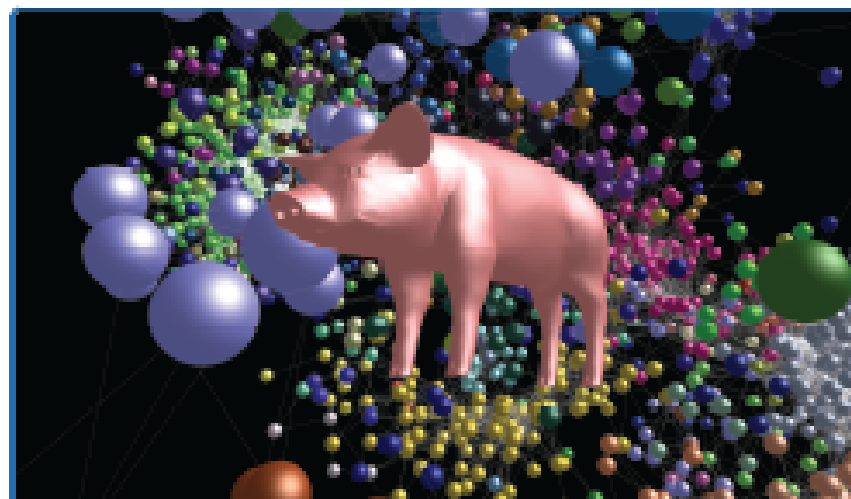
Genetic variation

- Several important 'mutations' missing
 - RYR1 (HAL), MSTN/GDF8 (double muscling), DGAT1 (milk yield)
- Indels missing
- Predictions - limitations



Gene expression

- From sequence to consequence
- From microarrays to RNAseq
- Expression atlases
 - Pig: microarray (published), RNAseq (in progress)
 - Sheep: RNAseq (in progress)
 - Chicken: RNAseq (partial)



A gene expression atlas of the domestic pig

Freeman *et al.*

Freeman *et al.* BMC Biology 2012, 10:90
<http://www.biomedcentral.com/1741-7007/10/90>

RESEARCH ARTICLE

Open Access

A gene expression atlas of the domestic pig

Tom C Freeman^{1*}, Alasdair Ivens^{2,6}, J Kenneth Baillie¹, Dario Beraldi^{1,7}, Mark W Barnett¹, David Dorward¹, Alison Downing¹, Lynsey Fairbairn¹, Ronan Kapetanovic¹, Sobia Raza¹, Andru Tomoiu¹, Ramiro Alberio³, Chunlei Wu⁴, Andrew I Su⁴, Kim M Summers¹, Christopher K Tuggle⁵, Alan L Archibald^{1*} and David A Hume^{1*}

Abstract

Background: This work describes the first genome-wide analysis of the transcriptional landscape of the pig. A new porcine Affymetrix expression array was designed in order to provide comprehensive coverage of the known pig transcriptome. The new array was used to generate a genome-wide expression atlas of pig tissues derived from 62 tissue/cell types. These data were subjected to network correlation analysis and clustering.

Results: The analysis presented here provides a detailed functional clustering of the pig transcriptome where transcripts are grouped according to their expression pattern, so one can infer the function of an uncharacterized gene from the company it keeps and the locations in which it is expressed. We describe the overall transcriptional signatures present in the tissue atlas, where possible assigning those signatures to specific cell populations or pathways. In particular, we discuss the expression signatures associated with the gastrointestinal tract, an organ that was sampled at 15 sites along its length and whose biology in the pig is similar to human. We identify sets of genes that define specialized cellular compartments and region-specific digestive functions. Finally, we performed a network analysis of the transcription factors expressed in the gastrointestinal tract and demonstrate how they subdivide into functional groups that may control cellular gastrointestinal development.

Conclusions: As an important livestock animal with a physiology that is more similar than mouse to man, we provide a major new resource for understanding gene expression with respect to the known physiology of mammalian tissues and cells. The data and analyses are available on the websites <http://biogps.org> and <http://www.macrophages.com/pig-atlas>.

Keywords: pig, porcine, *Sus scrofa*, microarray, transcriptome, transcription network, pathway, gastrointestinal tract

- Tool for monitoring gene expression
- Inferring function of unknowns
 - Inform genome annotation
- Comparative functional genomics
 - Is pig kidney more/less like human kidney than mouse kidney?
- Microarray-based atlas
- RNAseq atlas in progress

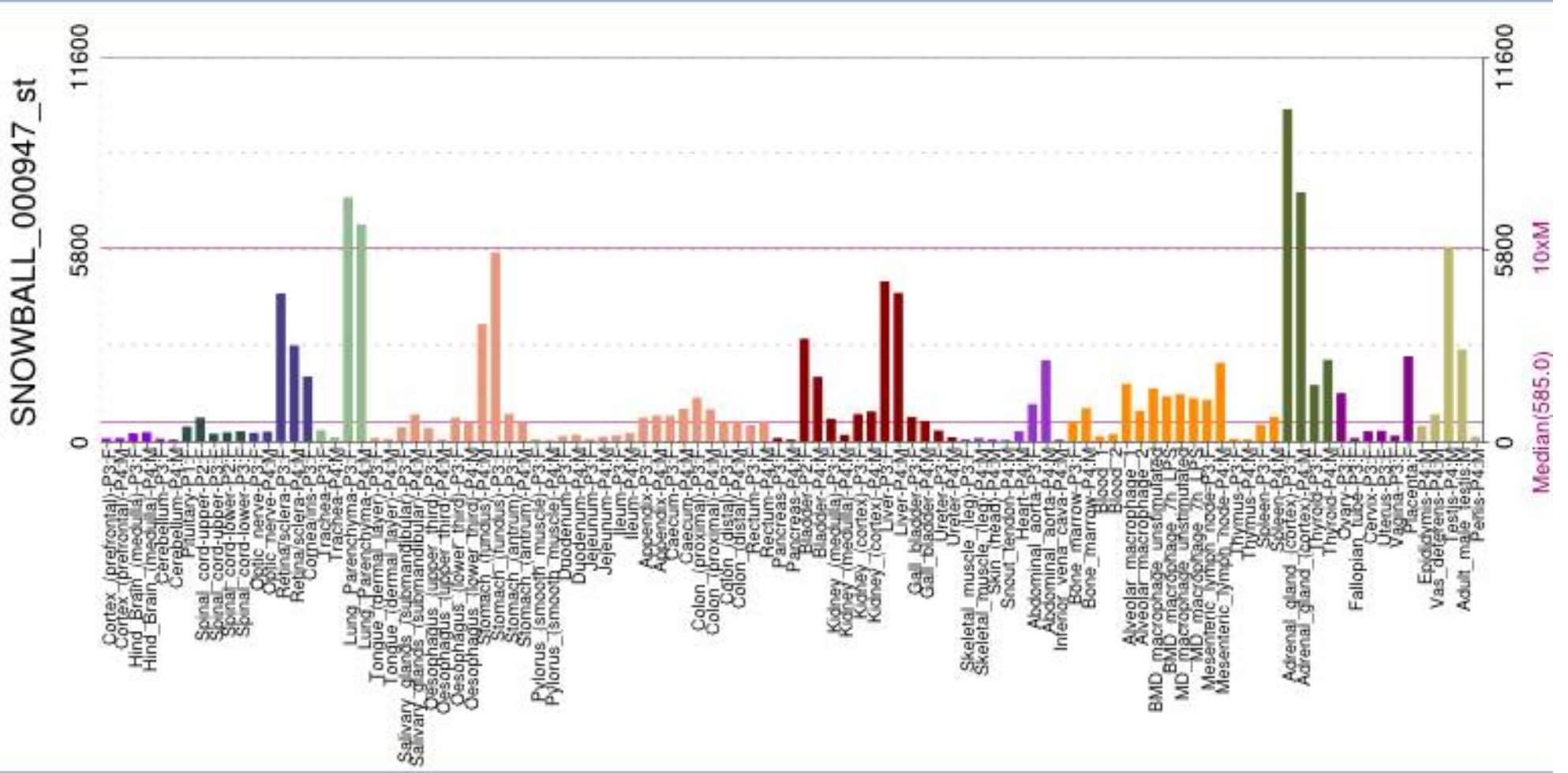
Affymetrix Porcine Snowball Array content



- 123 Affy controls
- 35 virus genomes (tiled 17 bp spacing)
- 1,857 miRNA probes
- 37 MT-mRNA
- 45,927 mRNA
 - 37,842 with annotation
 - 6,767 LOC annotations
 - 16,626 unique genes with official symbol/description



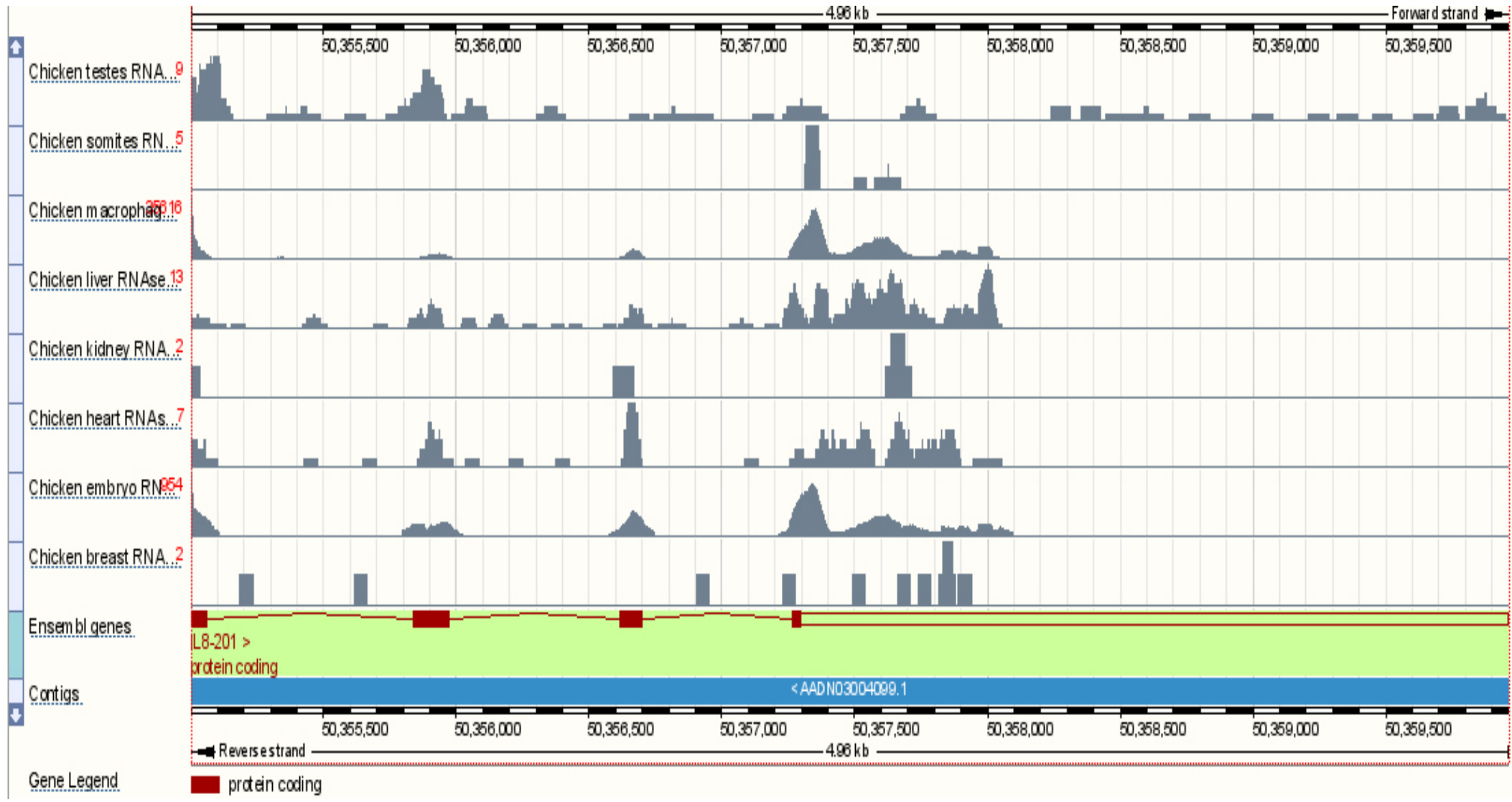
Expression profiles



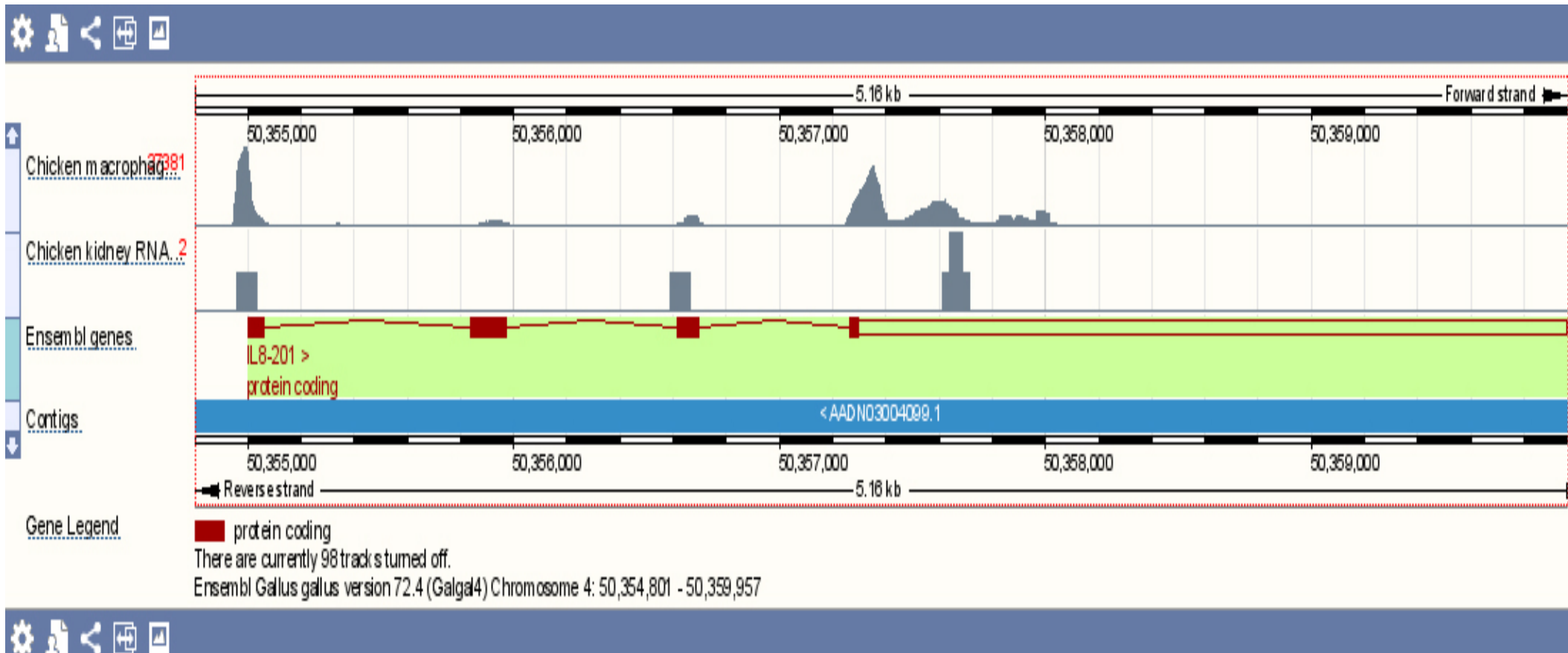
<http://biogps.org>



Chicken IL8 – RNAseq profiles



Chicken IL8 – RNAseq profiles



Sheep gene atlas

- Texel (ram, ewe, ewe lamb, 16d embryo)
- 50+ tissues per animal, whole embryo
 - Samples acquired, RNA prepared
- Illumina paired ends (2 x 150 bp)
 - > 1 Tb RNAseq data
- Ensembl RNAseq gene models (in progress)
- Funded 3SR, RoslinFoundation
- Poster 419



Cerebrum	Abomasum	Skeletal muscle, biceps	Testes, epididymis
Brain stem	Rumen	Skeletal muscle, longissimus dorsi	Corpus luteum, ovary, ovarian follicles
Tonsil	Duodenum	Skin (side/back)	Uterus, cervix, placenta
Cerebellum	Omentum	Spleen	Mammary gland
Hypothalamus	Caecum	Mesenteric lymph node	
Pituitary gland	Colon	Precapular lymph node	
Adrenal gland	Rectum	Peyer's patch	Whole embryo
Thyroid gland		Alveolar macrophages	
		Bone marrow	



Enabling the reading of farmed animal genomes

- Annotation of functional sequences
 - Protein coding
 - Non-coding RNA sequences
 - Regulatory sequences
- cf. human ENCODE project
 - Encyclopedia of DNA elements in the human genome

ENCODE



Encyclopedia of DNA Elements

Human

Data
Summary

Search

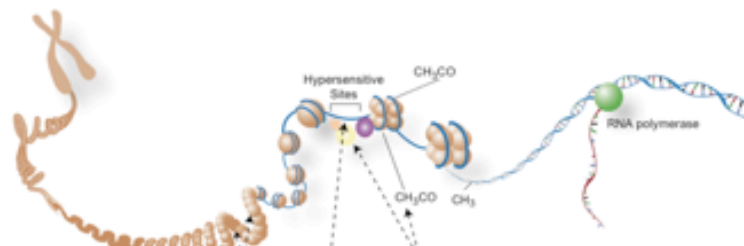
Downloads

Genome
Browser
(hg18)

Genome

About ENCODE Data

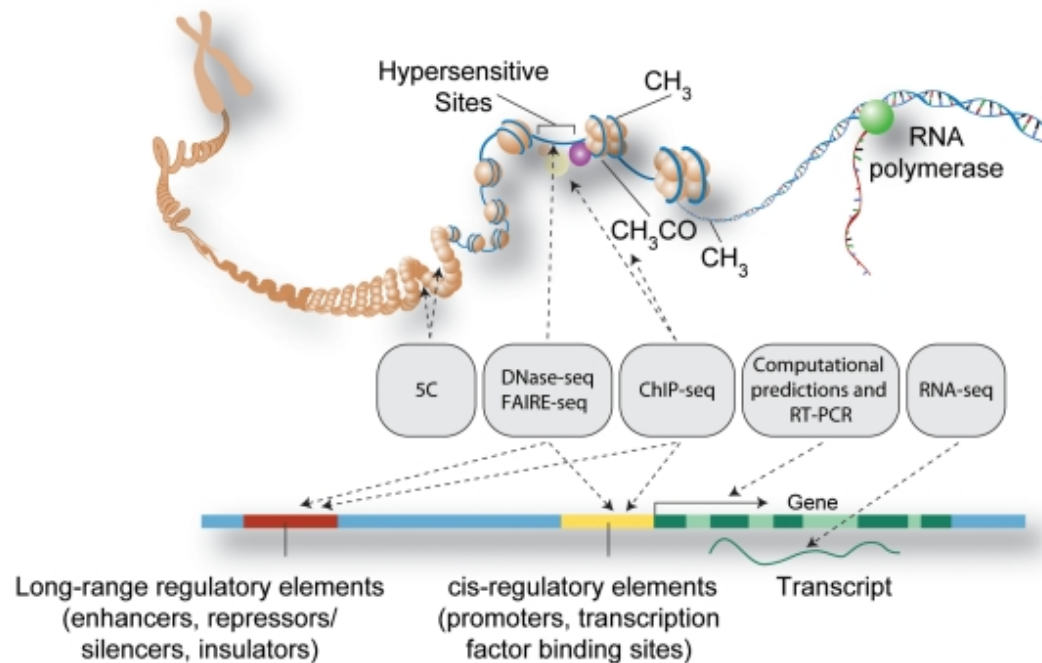
The [Encyclopedia of DNA Elements](#) (ENCODE) Consortium is an international collaboration of Research Institute ([NHGRI](#)). The goal of ENCODE is to build a comprehensive parts list of functions that act at the protein and RNA levels, and regulatory elements that control cells and circumstances.



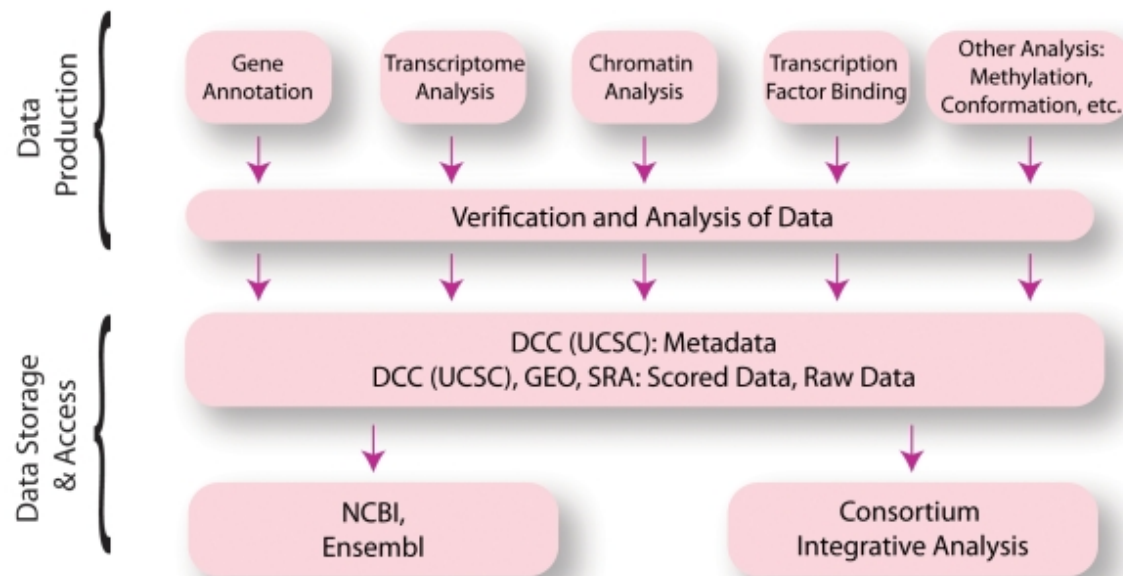
ENCODE data are now available for the entire genome. ***available for immediate use via :***

- [Search](#) for displayable tracks and download
- [Download](#) of data files

A.



B.

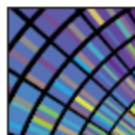


An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with



ENCODE
Encyclopedia of DNA Elements
nature.com/encode

95% of the genome lies within 8 kilobases (kb) of a DNA-protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.

Headlines

- It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters
- indicating that promoter functionality can explain most of the variation in RNA expression

Headlines

- Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes.
- In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.
 - See also Hindorff et al 2009 PNAS 106: 9362
 - 88% of trait associated SNPs are intronic / intergenic



ENCODE for farmed animals - Why?

- Understanding the genetic control of complex traits
- From sequence to consequence



ENCODE for farmed animals

- Genetic variation underlying trait variation
 - Coding sequence
 - RYR1, DGAT1,
 - Regulatory sequence
 - IGF2, callipyge
 - likely to be more important / common
- Current annotation limited
 - cDNA, EST-based gene models (now RNAseq models too)
 - SNP variation

ENCODE for farmed / companion animals

How

- By-product of biology-led research
 - development, differentiation, responses to perturbation
- Focus on target tissues
 - musco-skeletal
 - immune tissues
- Limited assays
 - DNaseI, FAIREseq
 - histone marks (promoters, enhancers)
 - methylation
 - RNAseq (**stranded**), CAGE

Table 2 | Summary of ENCODE histone modifications and variants

Histone modification or variant	Signal characteristics	Putative functions
H2A.Z	Peak	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Peak/region	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Peak	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Peak	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Peak	Mark of active regulatory elements with preference for promoters
H3K9me1	Region	Preference for the 5' end of genes
H3K9me3	Peak/region	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Peak	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Region	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Region	Elongation mark associated with transcribed portions of genes, with preference for 3' regions after intron 1
H3K79me2	Region	Transcription-associated mark, with preference for 5' end of genes
H4K20me1	Region	Preference for 5' end of genes

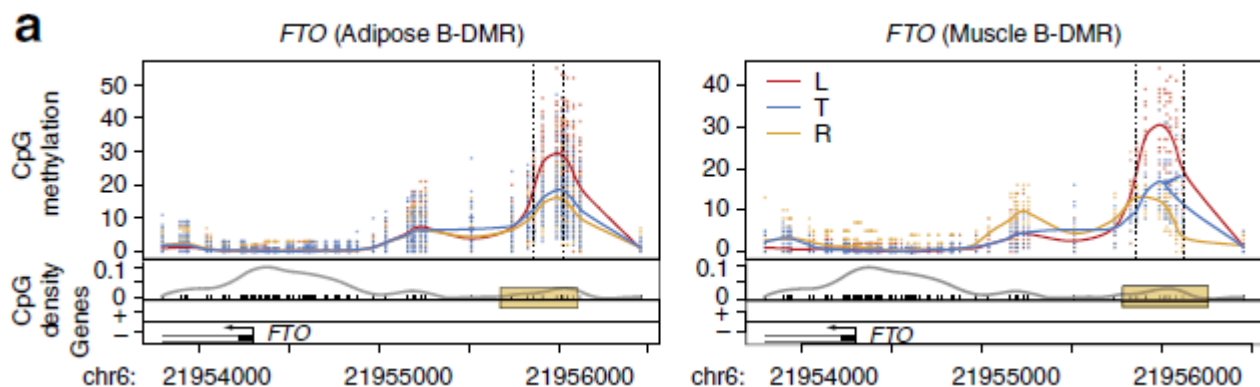
ARTICLE

Received 7 Oct 2011 | Accepted 19 Apr 2012 | Published 22 May 2012

DOI: 10.1038/ncomms1854



An atlas of DNA methylomes in porcine adipose and muscle tissues

Mingzhou Li^{1,*}, Honglong Wu^{2,*}, Zonggang Luo^{1,3}, Yudong Xia², Jiuqiang Guan¹, Tao Wang¹, Yiren Gu⁴, Lei Chen⁵, Kai Zhang^{1,†}, Jideng Ma¹, Yingkai Liu¹, Zhijun Zhong¹, Jing Nie¹, Shuling Zhou¹, Zhiping Mu¹, Xiaoyan Wang¹, Jingjing Qu¹, Long Jing¹, Huiyu Wang¹, Shujia Huang², Na Yi², Zhe Wang², Dongxing Xi², Juan Wang², Guangliang Yin², Li Wang², Ning Li², Zhi Jiang², Qiulei Lang⁶, Huasheng Xiao⁷, Anan Jiang¹, Li Zhu¹, Yanzhi Jiang¹, Guoqing Tang¹, Miaomiao Mai¹, Surong Shuai¹, Ning Li⁸, Kui Li⁹, Jinyong Wang⁵, Xiuqing Zhang², Yingrui Li², Haosi Chen¹⁰, Xiaolian Gao¹⁰, Graham S. Plastow¹¹, Stephen Beck¹², Huanming Yang², Jian Wang², Jun Wang², Xuwei Li¹ & Ruiqiang Li^{2,†}



Resource

Comparative Epigenomic Annotation of Regulatory DNA

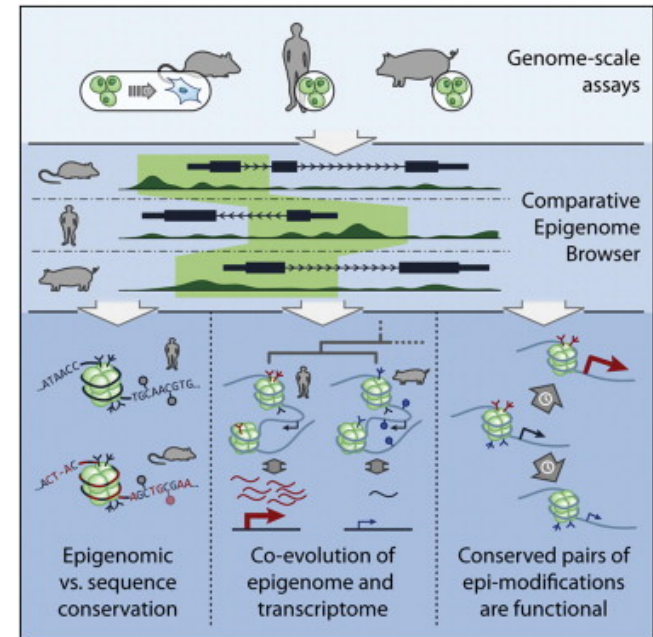
Shu Xiao^{1,2,6}, Dan Xie^{1,2,6}, Xiaoyi Cao^{2,3,6}, Pengfei Yu^{2,3,6}, Xiaoyun Xing⁵, Chieh-Chun Chen^{1,2}, Meagan Musselman¹, Mingchao Xie⁵, Franklin D. West⁴, Harris A. Lewin², Ting Wang⁵, Sheng Zhong^{1,2,3}  

¹ Department of Bioengineering, University of Illinois at Urbana-Champaign, 1304 West Springfield Avenue, Urbana, IL 61801, USA

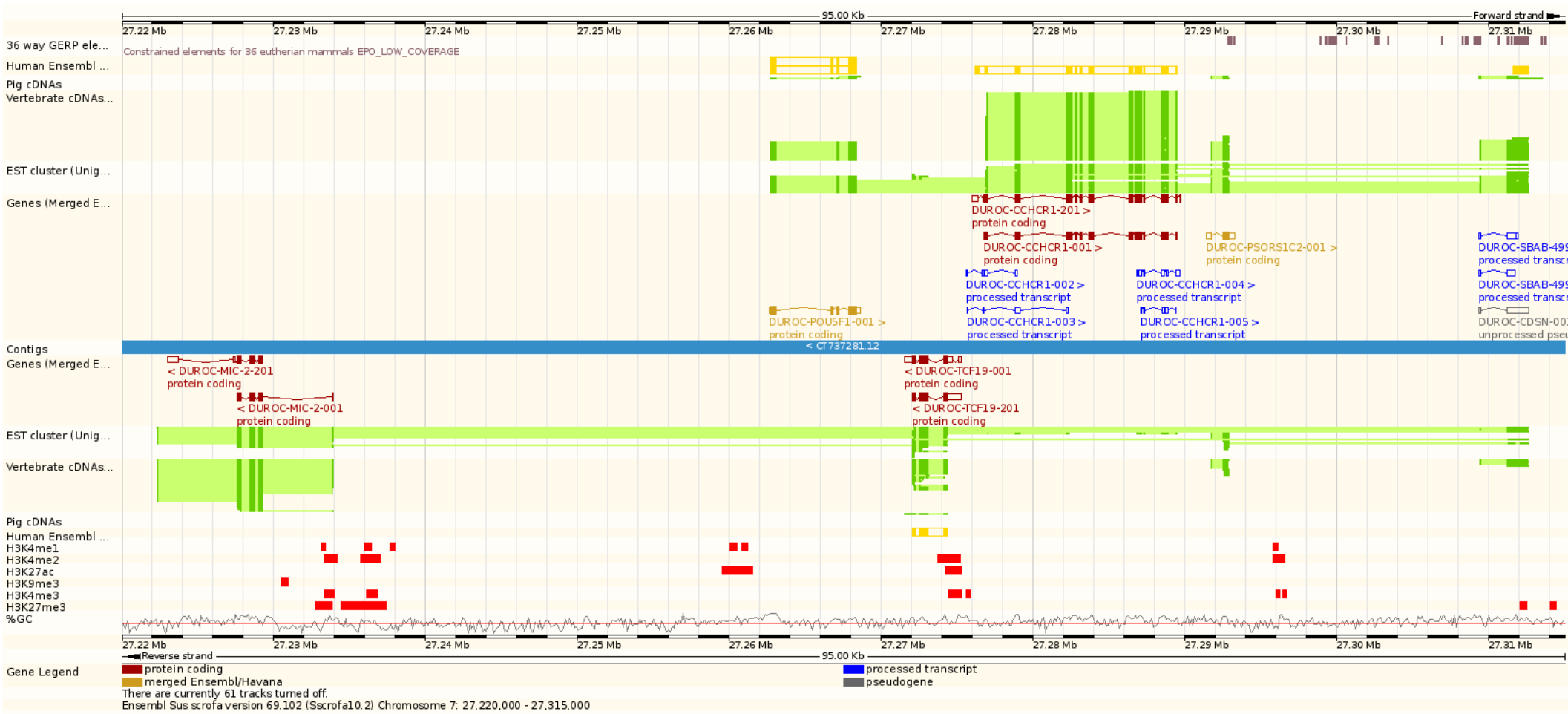
² Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1304 West Springfield Avenue, Urbana, IL 61801, USA

³ Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, 1304 West Springfield Avenue, Urbana, IL 61801, USA

⁴ Department of Animal and Dairy Science, University of Georgia, 425 River Road, Athens, GA 30602, USA







ENCODE for farmed / companion animals

- Species
 - single consortium / one per species / species groups
- Cells
 - transformed cells / primary cells / iPS cells
 - sharing
- Data management, publication
 - across groups
 - wider community
 - Data hubs model
 - Toronto Statement
- Coordination



A User's Guide to the Encyclopedia of DNA Elements (ENCODE)

The ENCODE Project Consortium^{†*}

Abstract

The mission of the Encyclopedia of DNA Elements (ENCODE) Project is to enable the scientific and medical communities to interpret the human genome sequence and apply it to understand human biology and improve health. The ENCODE Consortium is integrating multiple technologies and approaches in a collective effort to discover and define the functional elements encoded in the human genome, including genes, transcripts, and transcriptional regulatory regions, together with their attendant chromatin states and DNA methylation patterns. In the process, standards to ensure high-quality data have been implemented, and novel algorithms have been developed to facilitate analysis. Data and derived results are made available through a freely accessible database. Here we provide an overview of the project and the resources it is generating and illustrate the application of ENCODE data to interpret the human genome.



Data Standards

[Guidelines for Experiments](#)

The ENCODE Consortium has adopted uniform guidelines for the most common ENCODE experiments. The guidelines have evolved over time, as technologies have changed. The current guidelines are informed by results gathered during the project. Previous versions of the standards are also posted for reference.

[Platform Characterization](#)

ENCODE datasets are collected using a variety of platforms, such as ChIP-seq and RNA-seq. The consortium has undertaken several efforts to characterize these platforms to better understand the data being collected using them. Some of these efforts are described on this page.

[Quality Metrics](#) ^{New}

The ENCODE consortium analyzes the quality of the data produced using a variety of metrics. To view the quality metrics for many ENCODE datasets, go to this page. The quality metrics will be updated on occasion to include more recent data. Note that antibody validation information can be found at [ENCODE Antibodies](#). The **Platform Characterization** page, described above also examines an issue related to data quality.

Updated 04 June 2012

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, et al.

Genome Res. 2012 22: 1813-1831

Next steps

- White paper
- Develop data management strategy
- Review / promote ENCODE experimental protocols
- Develop / review cell line resources
- Develop communications strategy
- COST Action application (Sept 2013)
- US-EU ABWG workshop at PAG 2014



Where next?

- Improving reference genomes*
- Functional annotation (cf. ENCODE)*
- Sequencing 1000's to millions of individuals
 - Genotyping-by-sequencing and imputation
- Genomic selection



Acknowledgements

- Martien Groenen, Larry Schook, James Kijas, Jim Reecy
- Lel Eory, Steve Searle, Paul Flicek, Tim Hubbard
- Ewan Birney
- Jen Harrow, Jane Loveland



**edinburgh
genomics.**



**edinburgh
genomics.**

Coming soon!

Edinburgh Genomics will be a new facility formed from the merger of ARK Genomics and the GenePool, both world-class genomics facilities.

For further information, please e-mail us.

<http://genomics.ed.ac.uk>



- Merger of ARK-Genomics & Gene Pool
- Sequencing
 - Sanger: ABI3730
 - Illumina: 6x HiSeq2500, 3 MiSeq
- Bioinformatics
 - bioinformaticians
 - Edinburgh Parallel Computing Centre
 - secure multi terrabyte data store
 - secure compute Grid
- Genotyping, gene expression (ARK-Genomics)



DNA Sequencing

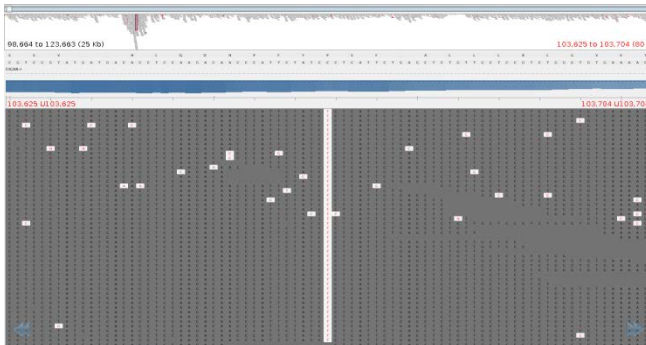
Illumina Sequencing

- Up to 250 bp paired
- Novel genomes
- Resequencing
- RNA-Seq
- ChIP-Seq
- Epigenetics

Illumina HiSeq 2500

Illumina MiSeq

Sanger 3730



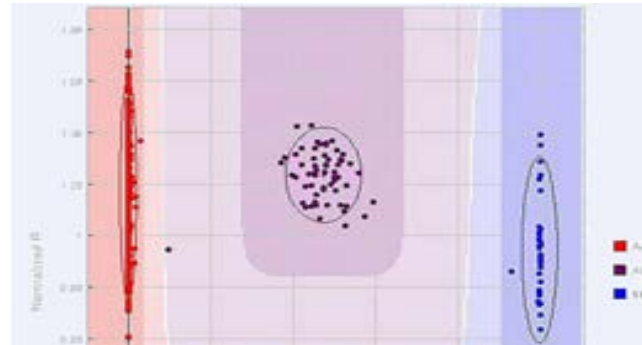
Genotyping

Illumina - from HD to custom chips

- iScan, Infinium
- BeadXpress, Goldengate
- BeadChip

Affymetrix

- GeneTitan, Axiom
- Process 96 arrays / run



Microarrays

Gene Expression

- Affymetrix
- Agilent
- Illumina
- Whole genome
- Exon-level
- microRNA

CGH, ChIP-Chip, MeIP

- Nimblegen

