Next Generation Sequencing – The Role of New Sequence Technologies in Shaping the Future of Veterinary Science

Hosted by the RCVS Charitable Trust

RCVS charitable trust

ROSLIN

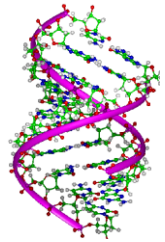
Sequencing animal genomes

Alan Archibald
The Roslin Institute and R(D)SVS
University of Edinburgh

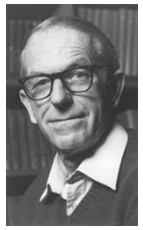
THE UNIVERSITY *of* EDINBURGH

BBSRC bioscience for the future


A sequenced genome is a requirement
for 21st Century biological research



1953
Watson and Crick



1977
DNA sequenced
ΦX174
5,386 nt

1990
Human Genome Project launched

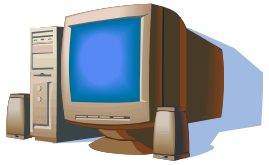
1991
PiGMaP project starts

'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

2002
Mouse draft genome sequence

2003
Human genome sequence "finished"
$3 billion

2004
Chicken genome sequenced

2005
Dog genome sequenced

2007
Cat genome sequenced

2008
Human 1000 Genomes Project launched

2009
Cattle genome sequenced

Horse genome Sequenced

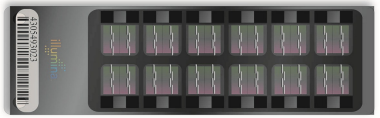
Mouse genome "finished"

2012
Pig genome sequenced
$20 million

2008
Bovine 50K SNP chip

2009
Pig 60K SNP chip ~$150

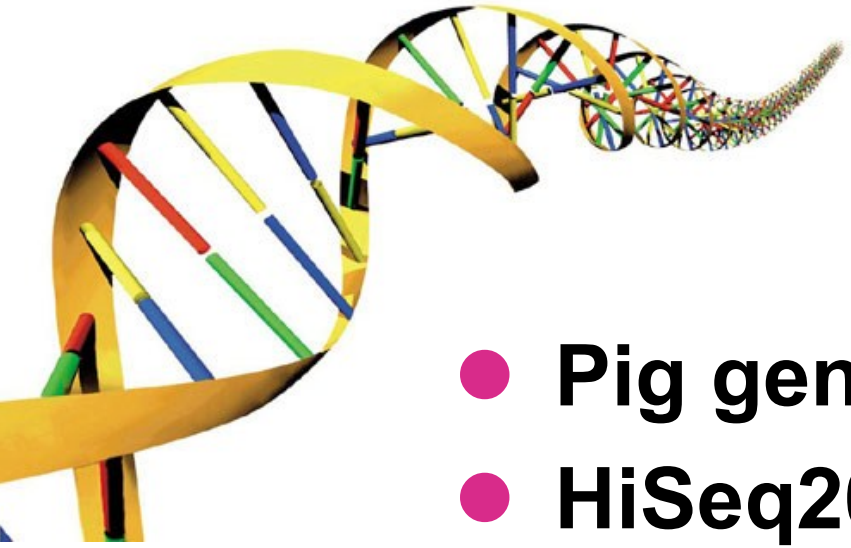
Sheep 60K SNP chip



2010
750K bovine SNP chips

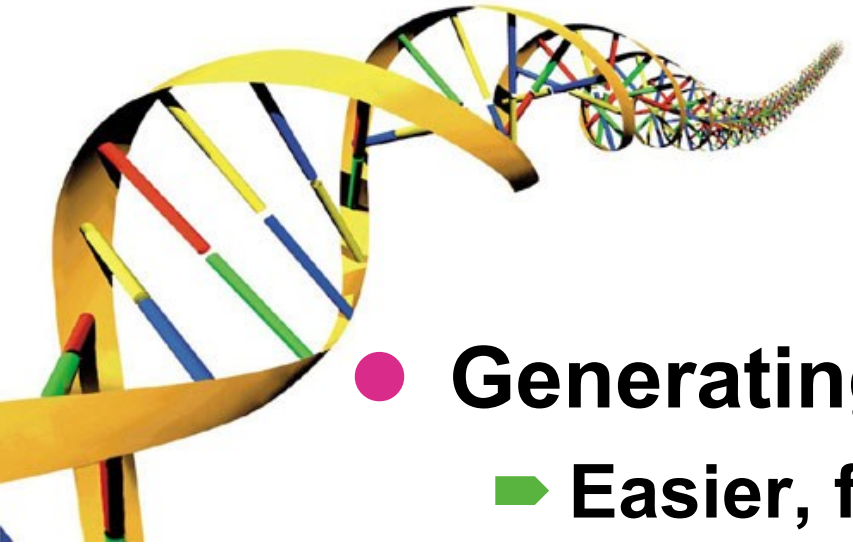
2012
Chicken 600K SNP chip

Goat 60K SNP chip



DNA sequencing

- Pig genome = ~2.7 Gb
- HiSeq2000 generates per run:
  - 600 Gb raw sequence (100 bp PE)
  - i.e. ~ 200x pig genome
  - in ~ 3×10^9 paired pieces
  - 70% expected to pass chastity filter
    - Q score > 20 for full 100 bp
- But . . . . .

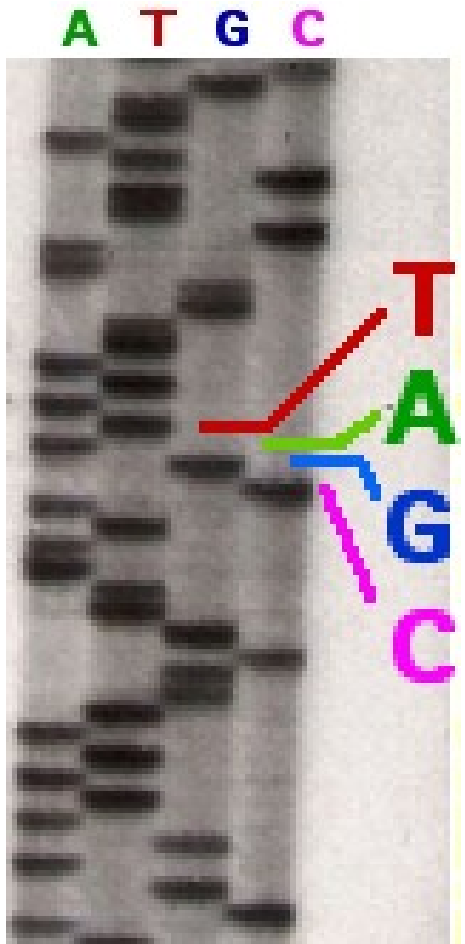


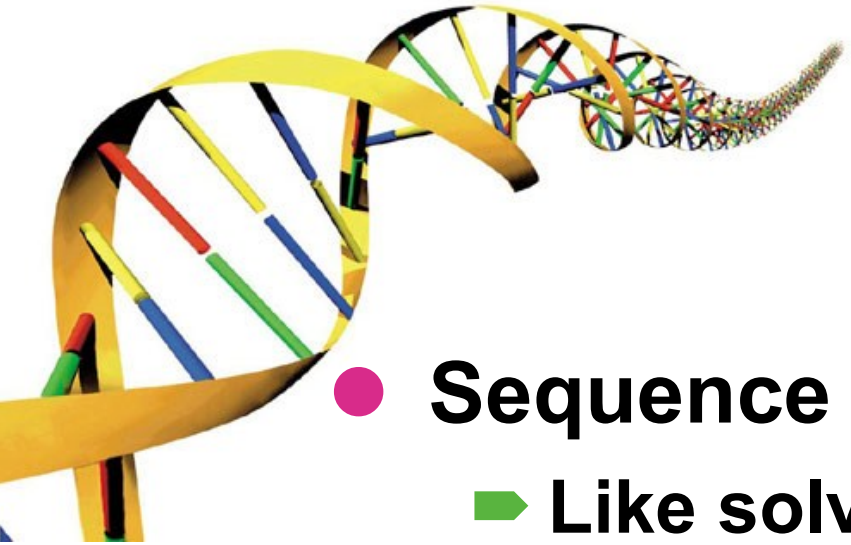
DNA sequencing

- Generating DNA sequence data is:
  - Easier, faster, cheaper
- All DNA sequencing technologies:
  - Generate errors
- Pig nucleus
  - Contains two haploid genomes
  - ~ 1 in 250 nt polymorphic
  - SNPs, indels, CNV

DNA sequencing

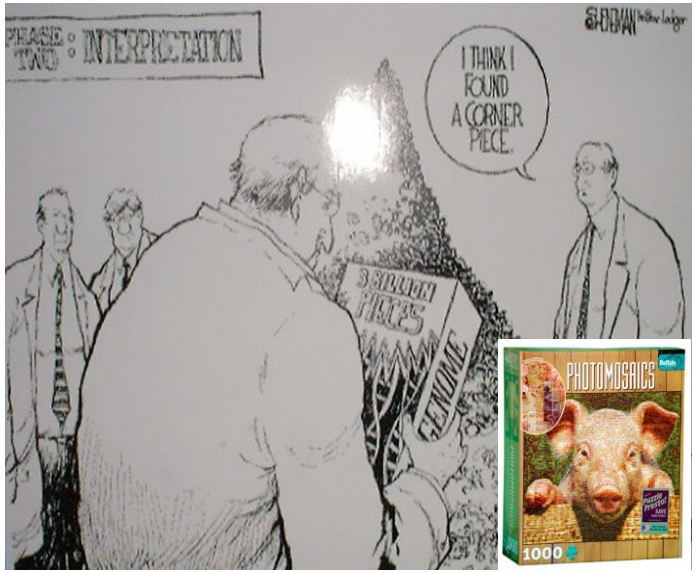
- Sequence multiple cloned molecules
  - Nested deletions
  - 2-300 bp reads
  - 5 kbp gene = PhD
- Shotgun approaches
  - ‘random’ fragmentation
  - high read depth





DNA sequencing

- **Sequence assembly is not:**
  - **Like solving a jigsaw puzzle**
  - **40% repetitive DNA sequence**
    - **cf. cloudless blue sky**
- **Building a high quality reference genome sequence takes:**
  - **Time, money, effort**



Pig Genome Sequencing Project

Swine Genome Sequencing Consortium

Comparative and Functional Genomics
Comp Funct Genom 2005; **6**: 251–255.
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/cfg.479

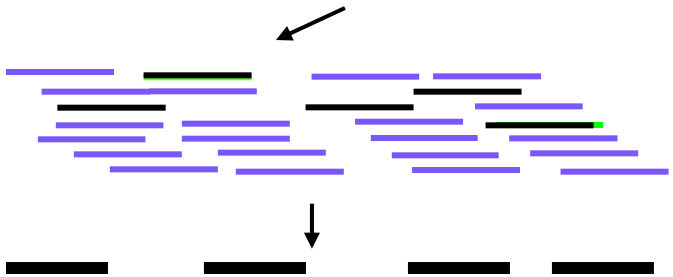
Conference Review

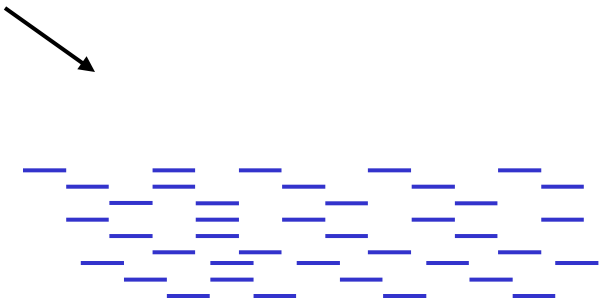
Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome

Lawrence B. Schook[1,2]*, Jonathan E. Beever[1,2], Jane Rogers[3], Sean Humphray[3], Alan Archibald[4], Patrick Chardon[5], Denis Milan[6], Gary Rohrer[7] and Kellye Eversole[8]

Hybrid Shotgun Sequencing Strategy

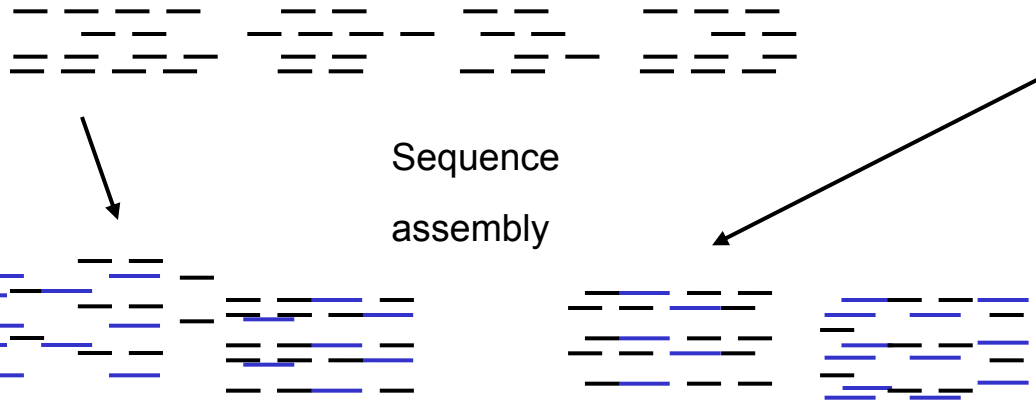
Minimal set of overlapping BACs selected from physical map





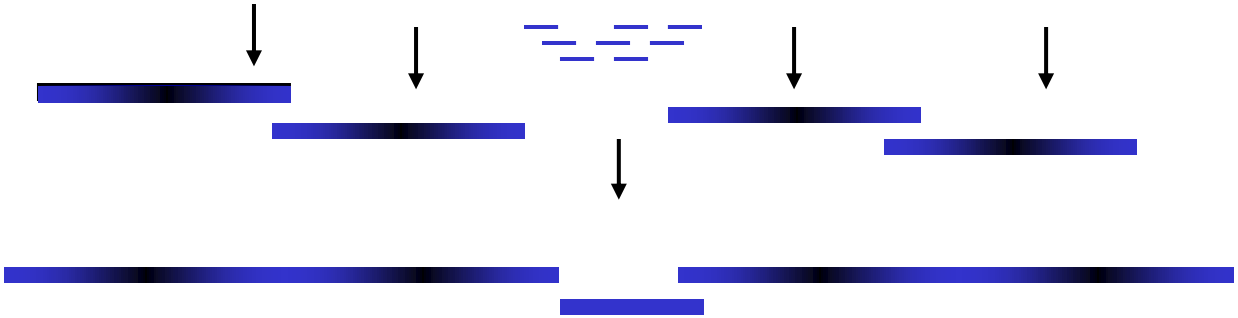
Whole- genome shotgun reads

BAC shotgun reads



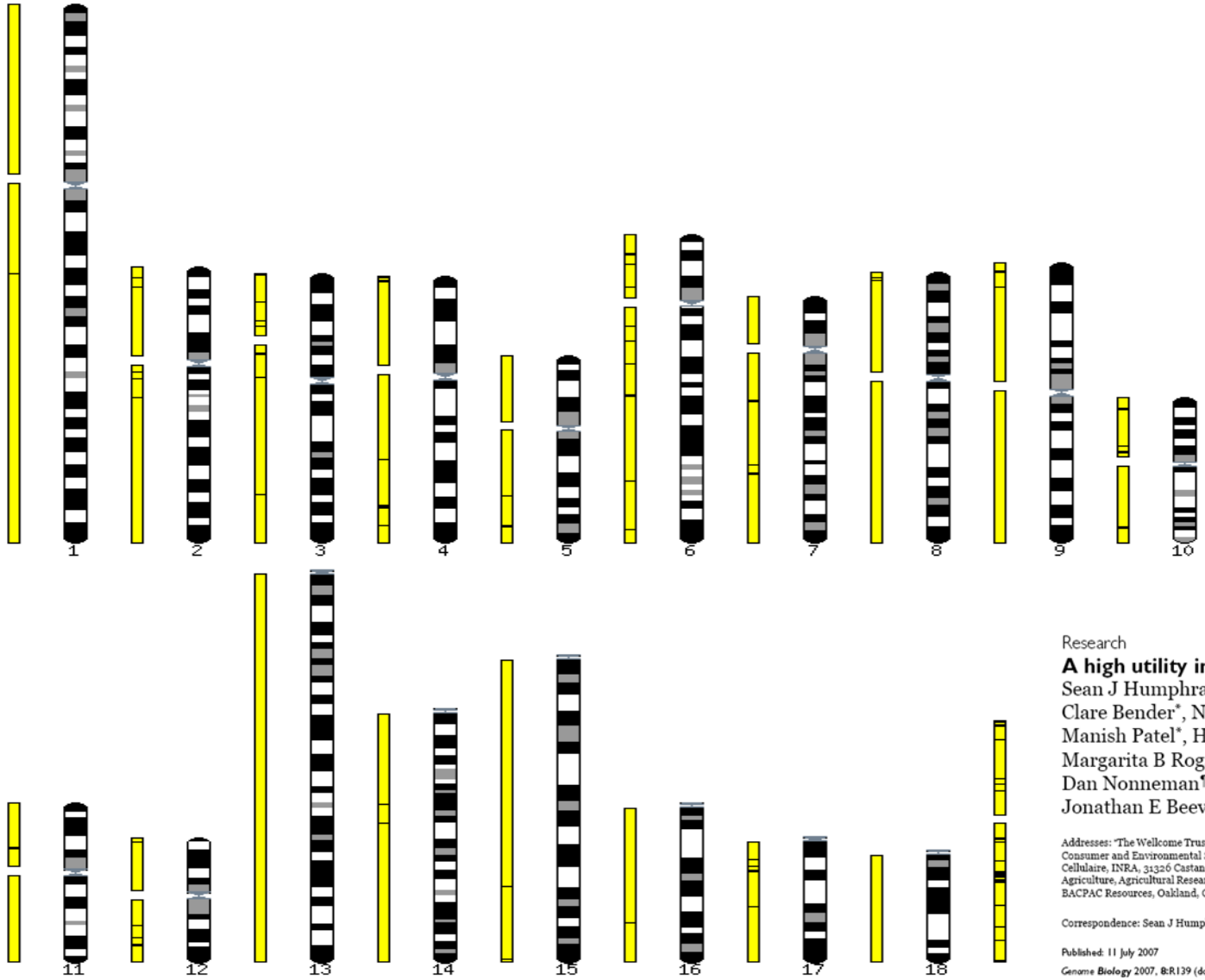
Sequence assembly

Combine overlapping whole-genome and BAC-derived reads



Assemble clone sequences to represent chromosomes and annotate using Ensembl automated pipeline

The best clone-based physcial map of a mammal



- 172 placed contigs
- average length 15 Mb
- covering 2.58Gb
- >98% of euchromatin

Humphray *et al.*, 2007. *Genome Biology*

Open Access

Research

A high utility integrated map of the pig genome

Sean J Humphray[*], Carol E Scott[*], Richard Clark[*], Brandy Marron[†], Clare Bender[*], Nick Camm[*], Jayne Davis[*], Andrew Jenks[*], Angela Noon[*], Manish Patel[*], Harminder Sehra[*], Fengtang Yang[*], Margarita B Rogatcheva[†], Denis Milan[‡], Patrick Chardon[§], Gary Rohrer[¶], Dan Nonneman[¶], Pieter de Jong[¥], Stacey N Meyers[†], Alan Archibald[#], Jonathan E Beever[†], Lawrence B Schook[†] and Jane Rogers[*]

Addresses: [*]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA UK. [†]College of Agriculture, Consumer and Environmental Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801 USA. [‡]Laboratoire de Génétique Cellulaire, INRA, 31326 Castanet-Tolosan, France. [§]INRA-CEA, Domaine de Vilvert, 78352, Jouy en Josas cedex, France. [¶]US Department of Agriculture, Agricultural Research Service, US Meat Animal Research Center, Clay Center, NE 68933-0166, USA. [¥]Children's Hospital Oakland-BACPAC Resources, Oakland, California 94609, USA. [#]Roslin Institute, Roslin, Midlothian EH25 9PS, UK.

Correspondence: Sean J Humphray. Email: sjh@sanger.ac.uk

Published: 11 July 2007

Genome Biology 2007, 8:R139 (doi:10.1186/gb-2007-8-7-r139)

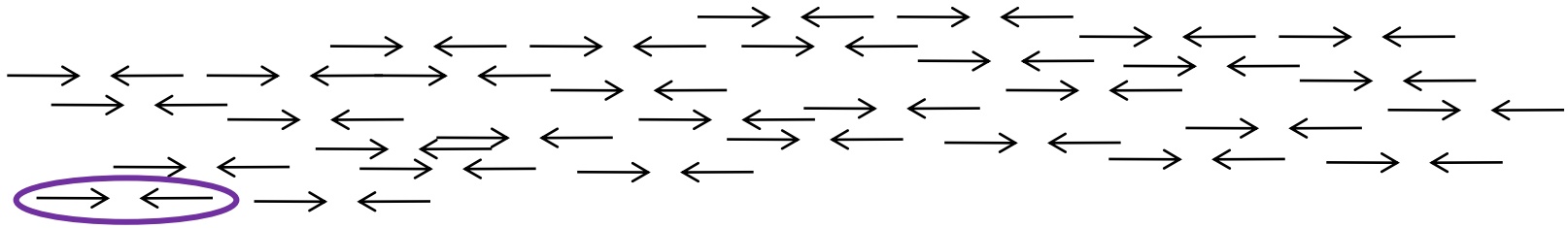
Received: 12 March 2007
Revised: 21 June 2007

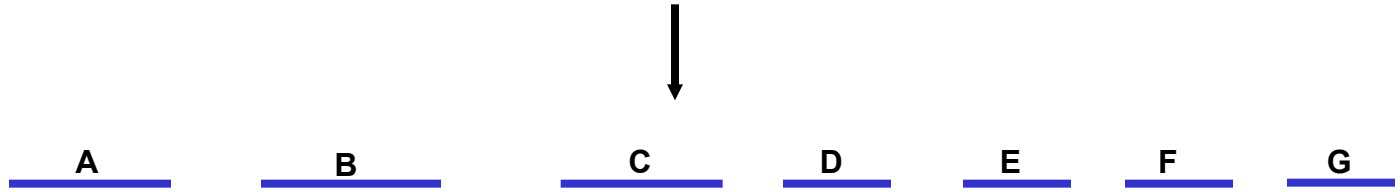
BAC Contigs / Fragments

end sequences
of subclone libraries

768 subclones / BAC
~3-4x coverage

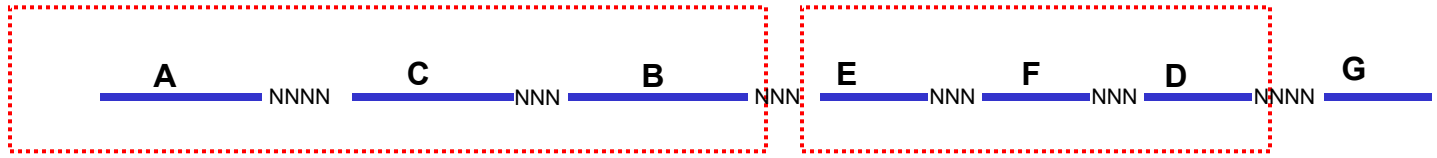
phrap





create fragment chains

Submission to EMBL/Genbank



fragment chain 1

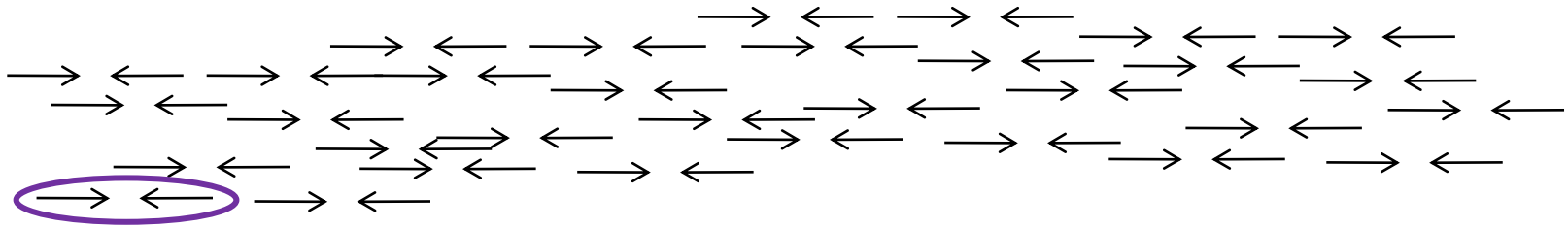
fragment chain 2

BAC Contigs / Fragments

end sequences
of subclone libraries

768 subclones / BAC
~3-4x coverage

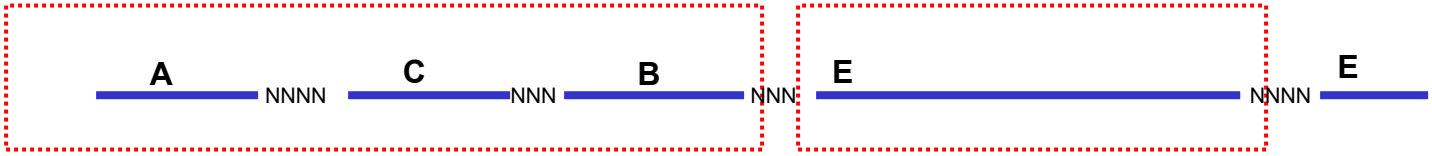
phrap



1 round automated primer
design and walk

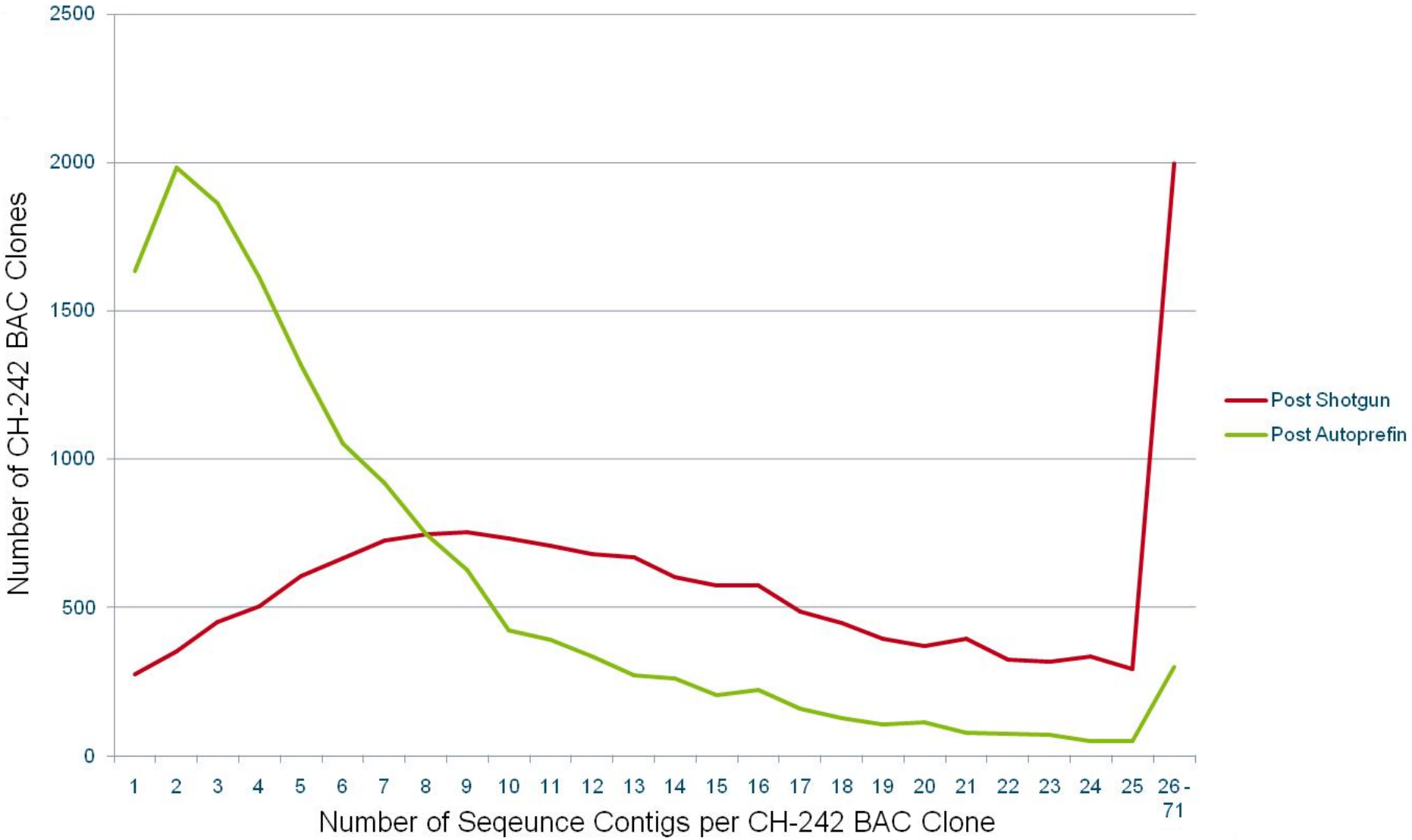
create fragment chains

Submission to EMBL/Genbank



fragment chain 1

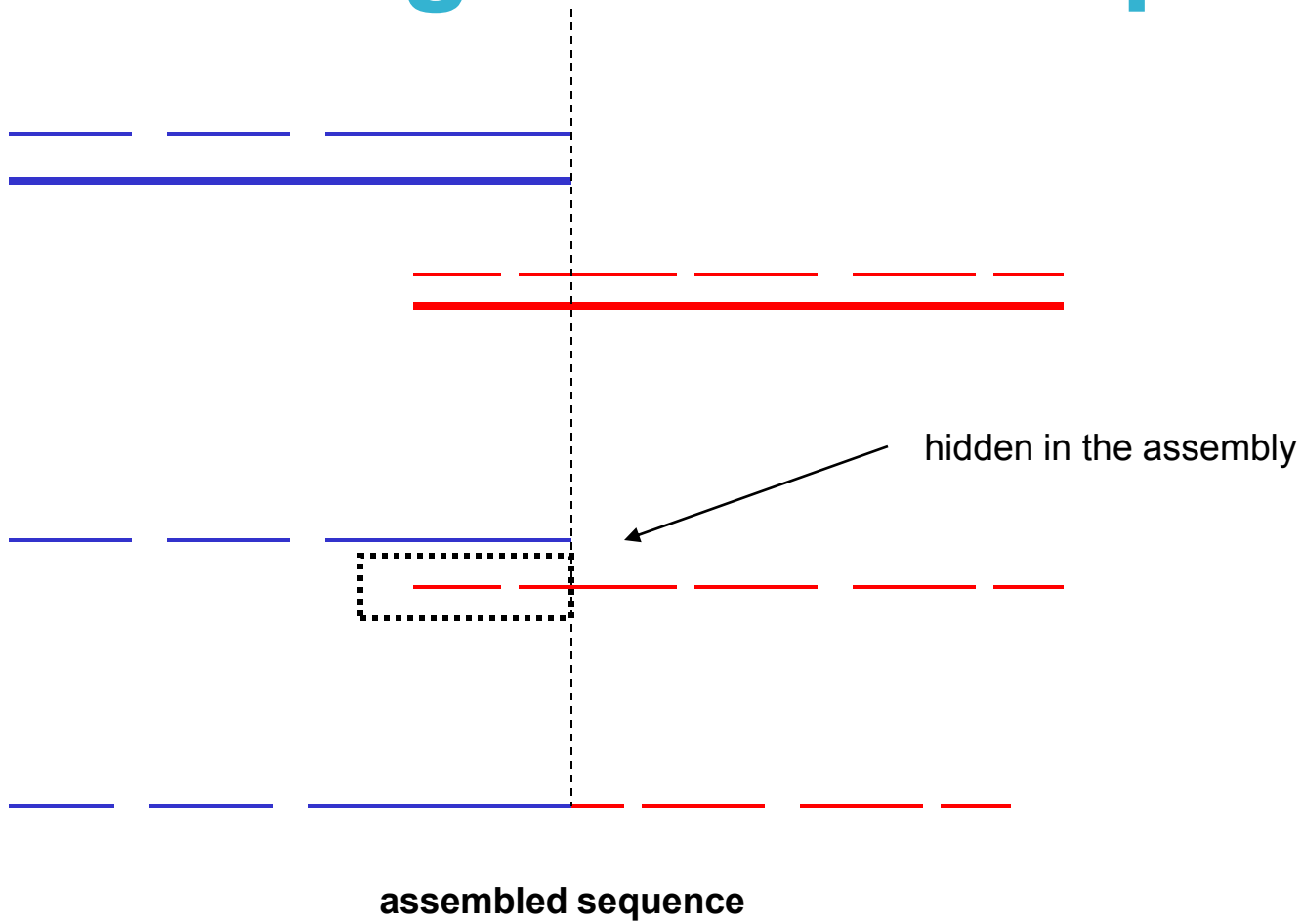
fragment chain 2


2500
2000
1500
1000
500
0
Number of CH-242 BAC Clones
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26-71
Number of Seqeunce Contigs per CH-242 BAC Clone
Post Shotgun
Post Autoprefin

Sequence improvement

	Sequence coverage	Sequence accuracy	Number of sequence contigs	Average contig length bp	Genes with incorrect structures
4x draft sequence	*97% (~80Mb missing)*	*99.99% accurate (1 error in10kbp)*	*160,000 (7k / chr)*	*17,000*	*30%*
Improved draft sequence	*99% (~32Mb missing)*	*99.99% accurate (1 error in10kbp)*	*65,000 (3k / chr)*	*42,000*	*5%*
Gold standard finished sequence	*99.9% (~5Mb missing)*	*99.999% (1 error in 100kbp)*	*<200*	*14,000,000*	*0%*

Resolving BAC overlaps



Add WGS data

Same Duroc individual as CHORI-242

BGI

- **66.5 Gb of sequence (24-fold)**
- **Read length: 44**

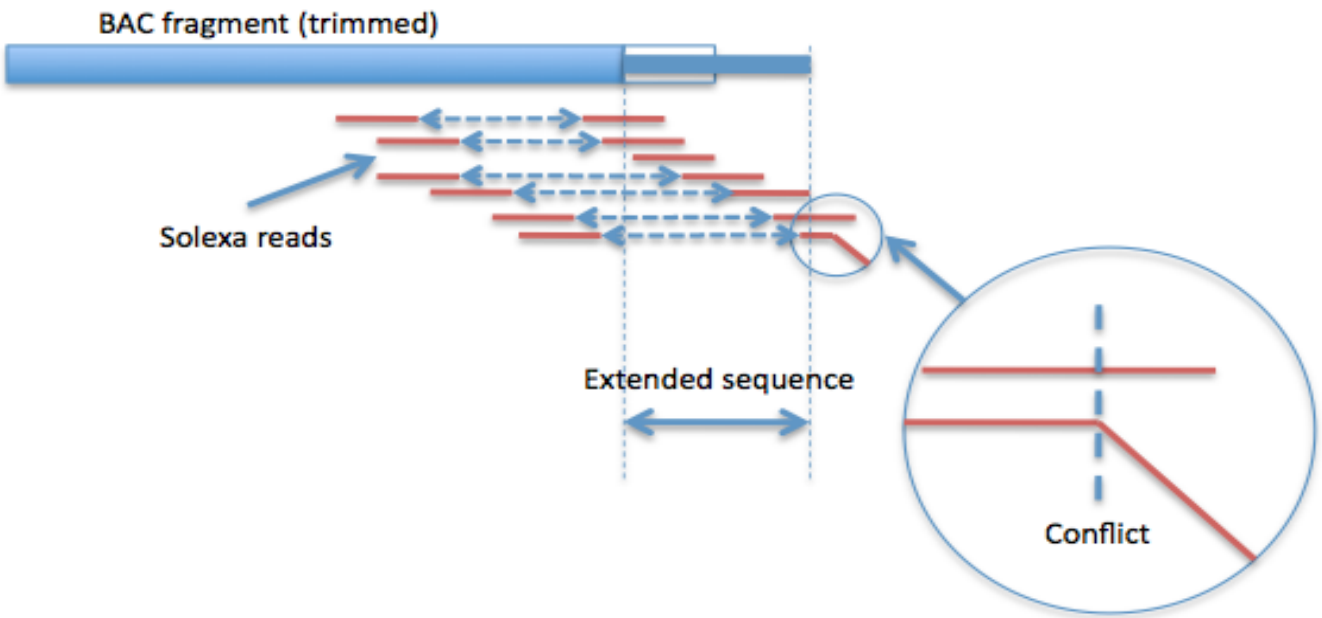
WTSI

- **~40 Gb of sequence (14-fold)**
- **Read length: 108**

Assemblies

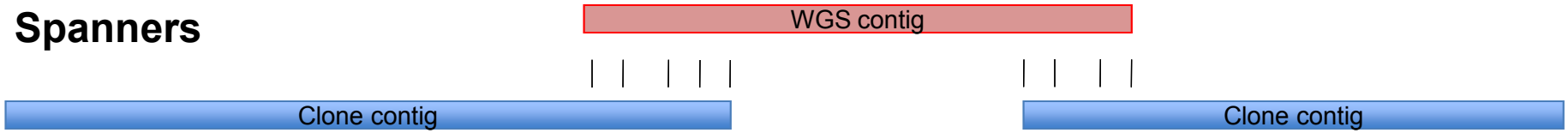
- **SOAPdeNovo (Shenting Li – Pig Genome III - Hinxton)**
  - **18,409 gaps closed (using reads)**
  - **31,359 gaps closed (using scaffolds)**
  - **252Mb in new scaffolds (BRCA1 gene)**
- **Cortex (Caccamo & Iqbal)**
  - **~3000 gaps closed – use to correct issues within BAC assemblies**

Illumina Assembly – using reads

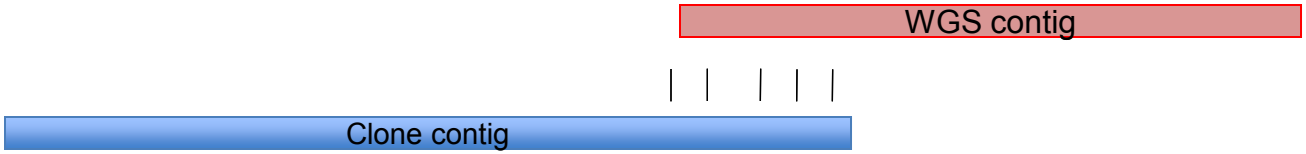


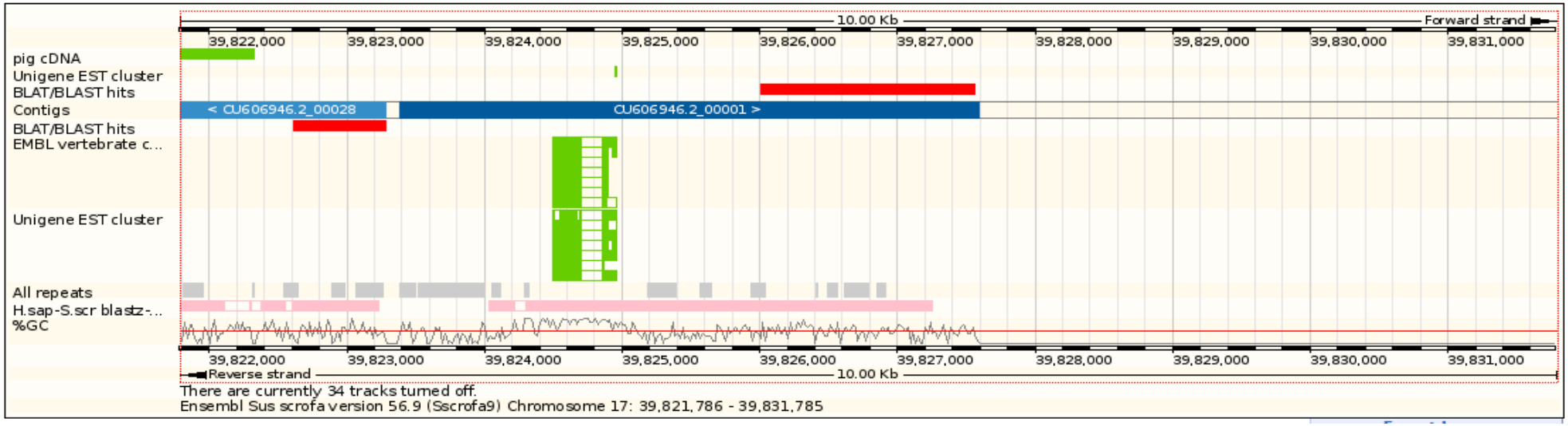
Illumina Assembly – using contigs

Spanners

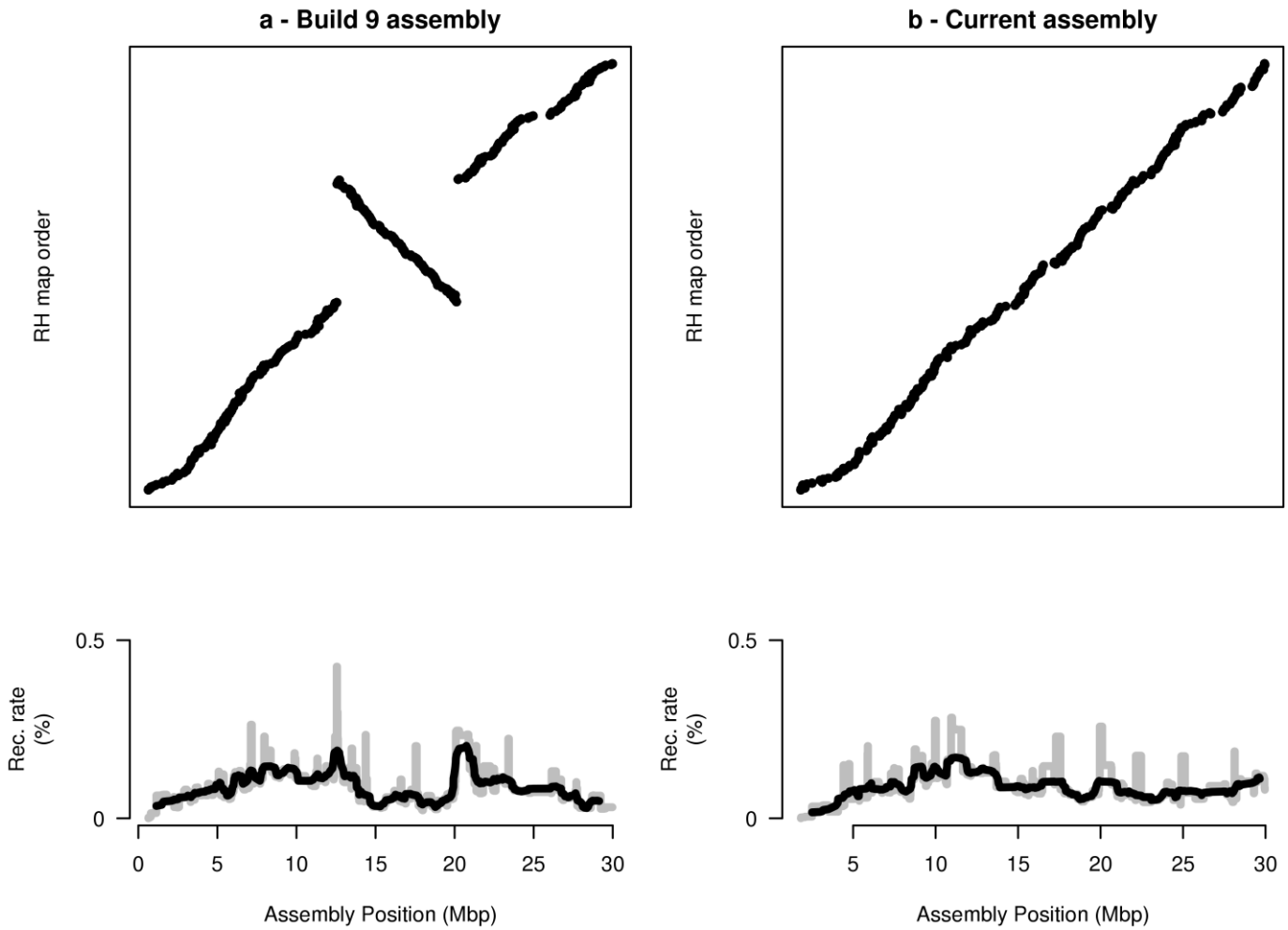


Hangers




10.00 Kb
Forward strand
39,822,000
39,823,000
39,824,000
39,825,000
39,826,000
39,827,000
39,828,000
39,829,000
39,830,000
39,831,000
pig cDNA
Unigene EST cluster
BLAT/BLAST hits
Contigs
< CU606946.2_00028
CU606946.2_00001 >
BLAT/BLAST hits
EMBL vertebrate c...
Unigene EST cluster
All repeats
H.sap-S.scr blastz-...
%GC
Reverse strand
There are currently 34 tracks turned off.
Ensembl Sus scrofa version 56.9 (Sscrofa9) Chromosome 17: 39,821,786 - 39,831,785

Order-orientation Sequence vs RH maps



Annotated assembly in Ensembl

Login · Register

BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Pig (Sscrofa10.2)

About this species
- Description
- Genome Statistics
  - Assembly and Genebuild
  - Top 40 InterPro hits
  - Top 500 InterPro hits
- What's New
- Sample entry points
  - Karyotype
  - Location (7:60107914-60305
  - Gene (ENSSSCG000000042
  - Transcript (ENSSSCT00000
  - Variation (rs10719553)

Configure this page

Manage your data

Export data

Bookmark this page

Pig (Sus scrofa)

Search for: Go

e.g. ENSSSCG00000004244 or 7:60107914-60305245 or apoptosis

Description

Assembly

The Sscrofa10.2 assembly of the pig genome was produced in August 2011 by the Swine Genome Sequencing Consortium (SGSC). It consists of 20 chromosomes (1-18, X and Y) and 4562 unplaced scaffolds. This genome assembly has GCA_000003025.4 as its GenBank assembly accession.

The genome assembly represented here corresponds to GenBank Assembly ID GCA_000003025.4

Download Pig genome sequence (FASTA)

Previous assemblies

Sscrofa9 (Release 66, Feb2012) Go to archive

Annotation

Sscrofa10.2 was annotated using a standard Ensembl mammalian genebuild pipeline, incorporating RNA-Seq data provided by the (SGSC). The annotation process is described in this document. When the annotation was completed, the gene set contained the following: 21,640 protein coding genes, 380 pseudogenes and 2,965 ncRNAs.

- Detailed information on genebuild (PDF)

Ensembl release 67 - May 2012 © WTSI / EBI

About Ensembl | Privacy Policy | Contact Us

Permanent link - View in archive site

Done Internet 100%

http://www.ensembl.org/Sus_scrofa/Info/Index

Sscrofa10.2 – assembly, genes

Assembly			Annotation*	
	Placed	Unplaced		
Total length	2,596,639,456	211,869,922	Protein-coding genes:	21,627
Ungapped length	2,323,671,356	195,490,322	Pseudogenes:	380
Scaffolds	5,343	4,562	ncRNA genes**:	2,965
Contigs	73,524	168,358	Gene exons:	197,675
Scaffold N50	637,332	98,022	Gene transcripts:	26,487
Contig N50	80,720	2,423		

Data sharing and publication

- **Bermuda agreement**
- **Fort Lauderdale agreement**

nature Vol 461|10 September 2009

OPINION

Prepublication data sharing

Rapid release of prepublication data has served the field of genomics well. Attendees at a workshop in Toronto recommend extending the practice to other biological data sets.

Pig genomes provide insight into porcine demography, domestication and evolution

Martien A.M. Groenen[1*], Alan L. Archibald[2*], ***Hirohide Uenishi[3], Christopher K. Tuggle[4], Yasu Takeuchi[5], Max F. Rothschild[4], Claire Rogel-Gaillard[6], Chankyu Park[7], Denis Milan[8], Hendrik-Jan Megens[1], Shengting Li[9], Denis Larkin[10], Heebal Kim[11], Laurent A. F. Frantz[1], Mario Caccamo[12]***, Hyeonju Ahn[11], Bronwen L. Aken[13], Anna Anselmo[14], Christian Anthon[15], Loretta Auvil[16], Bouabid Badaoui[14], Craig W. Beattie[17], Christian Bendixen[18], Daniel Berman[19], Frank Blecha[20], Jonas Blomberg[21], Lars Bolund[9], Mirte Bosse[1], Sara Botti[14], Zhan Bujie[18], Megan Bystrom[4], Boris Capitanu[16], Denise Carvalho-Silva[22], Patrick Chardon[6], Celine Chen[24], Ryan Cheng[4], Sang-Haeng Choi[25], William Chow[13], Richard C. Clark[13], Christopher Clee[13], Richard P.M.A. Crooijmans[1], Harry D. Dawson[24], Patrice Dehais[8], Floravante De Sapio[2], Bert Dibbits[1], Nizar Drou[12], Zhi-Qiang Du[4], Kellye Eversole[26], João Fadista[18], Susan Fairley[13], Thomas Faraut[8], Geoffrey J. Faulkner[2], Katie E. Fowler[27], Merete Fredholm[15], Eric Fritz[4], James G.R. Gilbert[13], Elisabetta Giuffra[14], Jan Gorodkin[15], Darren K. Griffin[27], Jennifer L. Harrow[13], Alexander Hayward[28], Kerstin Howe[13], Zhi-Liang Hu[4], Sean J. Humphray[13], Toby Hunt[13], Henrik H. Jensen[18], Patric Jern[28], Matthew Jones[13], Jerzy Jurka[29], Hiroyuki Kanamori[30], Ronan Kapetanovic[2], Jaebum Kim[31,23], Jae-Hwan Kim[32], Kyu-Won Kim[33], Tae-Hun Kim[34], Greger Larson[35], Kyooyeol Lee[7], Kyung-Tai Lee[34], Richard Leggett[12], Harris A. Lewin[36], Yingrui Li[9], Wansheng Liu[37], Jane E. Loveland[13], Yao Lu[9], Joan K. Lunney[19], Jian Ma[38], Ole Madsen[1], Katherine Mann[19], Lucy Matthews[13], Stuart McLaren[13], Takeya Morozumi[30], Michael Murtaugh[39], Jitendra Narayan[10], Dinh Truong Nguyen[7], Peixiang Ni[9], Song-Jung Oh[40], Suneel Onteru[4], Frank Panitz[18], Eung-Woo Park[34], Hong-Seog Park[25], Geraldine Pascal[41], Yogesh Paudel[1], Miguel Perez-Enciso[42], Ricardo Ramirez-Gonzalez[12], James M. Reecy[4], Sandra Rodriguez-Zas[43], Gary A. Rohrer[44], Lauretta Rund[43], Yongming Sang[20], Kyle Schachtschneider[43], Joshua Schraiber[45], John Schwartz[39], Linda Scobie[46], Carol Scott[13], Stephen Searle[13], Bertrand Servin[8], Bruce R. Southey[43], Goran Sperber[47], Peter Stadler[48], Jonathan Sweedler[49], Hakim Tafer[48], Bo Thomsen[18], Rashmi Wali[46], Jian Wang[9], **Jun Wang[9]**, Simon White[13], Xun Xu[9], Martine Yerle[8], Jianguo Zhang[9], Guojie Zhang[9], Jie Zhang[50], Shuhong Zhao[50], **Jane Rogers[12], Carol Churcher[13] and Lawrence B. Schook[51]**.

Pig Genome Sequence Project:
A Blueprint for Agriculture, Life and Biomedical Sciences

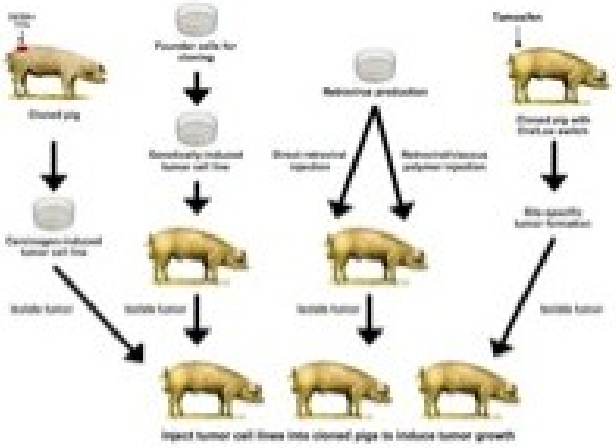
CGCTCATATCGATCGATTGCATCGATCCGATCGATCGATGCTAGCTAGCGAGTCGAA

DNA-Based Models

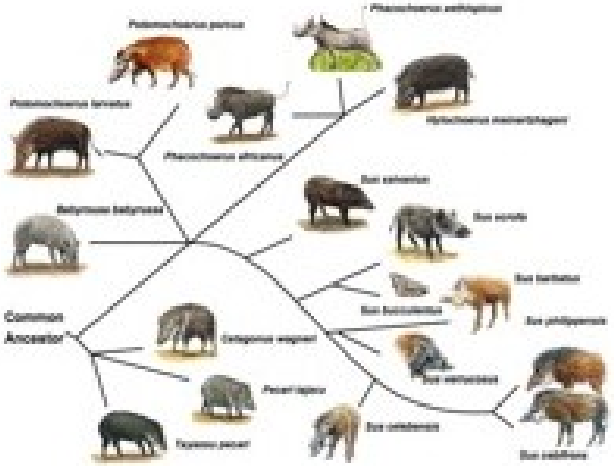
GGCTCTATCTATCCTTAACGTAGCTATATCGCTAGGCTAGTCCTGATCTGAGATCGA

TCAGGGCCATTATCGGCATCGATCGATTGATCAACCTGTAGAGCCATTGCTATTCGA

AGRICULTURAL SCIENCES
Genotypes and Phenotypes



BIOMEDICAL SCIENCES
Designing Cancer Models



LIFE SCIENCES
Mechanisms of Mammalian Evolution and Diversity

Acknowledgements - funding

Funding Source	Funding Level	Activity Supported
CSREES USDA	**$10,000,000**	**Clone shotgun sequencing**
USDA – ARS	**$1,000,000**	**Clone shotgun sequencing**
N Carolina Agric Res Service	**$50,000**	**Sequencing**
Iowa State University	**$200,000**	**Targeted sequencing**
USA National Pork Board	**$750,000**	**WGS sequencing**
Iowa Pork Producers Assn.	**$100,000**	**Targeted Sequencing**
N. Carolina Pork Council	**$100,000**	**Sequencing**
EU SABRE Project	**€1,600,000**	**6X of SSC7 & SSC14**
Danish government/BGI	***€600,000***	**Next-gen seq WGS**
INRA Genescope, France	**€1,000,000**	**SNP discovery (1 million reads)**
BBSRC, UK	***£2,000,000***	**Annotation, analysis; SSCX/Y**
Wellcome Trust Sanger Institute	**£1,600,000**	**Contribution to read costs**
Wellcome Trust Sanger Institute	**£300,000**	**Finish ENCODE + MHC**
Dutch IPG	**€600,000**	**6X of SSC4**
Japan	**£10,000**	**35 SSC7 clones**
Korea	**0.5 million reads**	**WGS sequencing**

Genomics-enabled tools

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

HIGH DENSITY SNP GENOTYPING CHIP

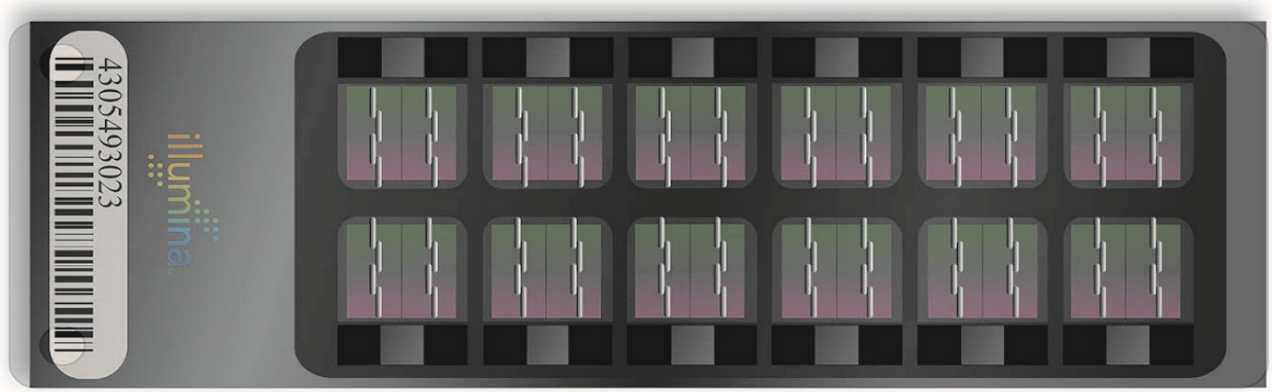
OPEN ACCESS Freely available online

August 2009 PLoS one

Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology

Antonio M. Ramos[1], Richard P. M. A. Crooijmans[1], Nabeel A. Affara[2], Andreia J. Amaral[1], Alan L. Archibald[3], Jonathan E. Beever[4], Christian Bendixen[5], Carol Churcher[6], Richard Clark[6], Patrick Dehais[7], Mark S. Hansen[8], Jakob Hedegaard[5], Zhi-Liang Hu[9], Hindrik H. Kerstens[1], Andy S. Law[3], Hendrik-Jan Megens[1], Denis Milan[7], Danny J. Nonneman[10], Gary A. Rohrer[10], Max F. Rothschild[9], Tim P. L. Smith[10], Robert D. Schnabel[11], Curt P. Van Tassell[12], Jeremy F. Taylor[11], Ralph T. Wiedmann[10], Lawrence B. Schook[4], Martien A. M. Groenen[1]*

1 Wageningen University, Animal Breeding and Genor
Kingdom, **3** Division of Genetics and Genomics, The R
Kingdom, **4** Institute for Genomic Biology, University
Denmark, **6** The Wellcome Trust Sanger Institute, Tl
Cellulaire, Castanet Tolosan, France, **8** Illumina, Inc., S



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
  - £1,000 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

Gene Expression Microarray

Gene Expression Atlas of the Pig

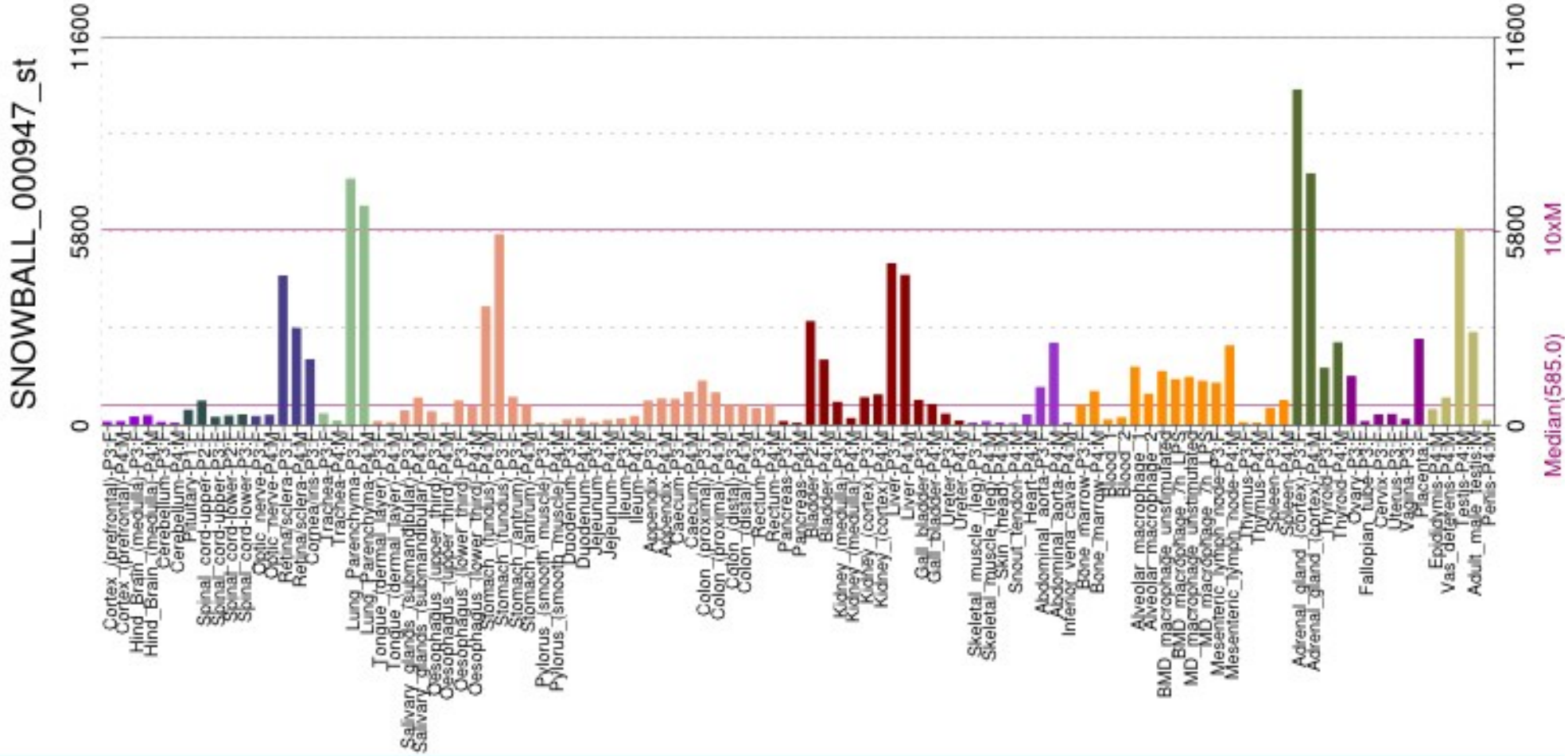
Expression array, atlas

- 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?
- Current arrays
  - Poorly annotated
  - Content elderly (e.g. Affymetrix 2004 design)

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

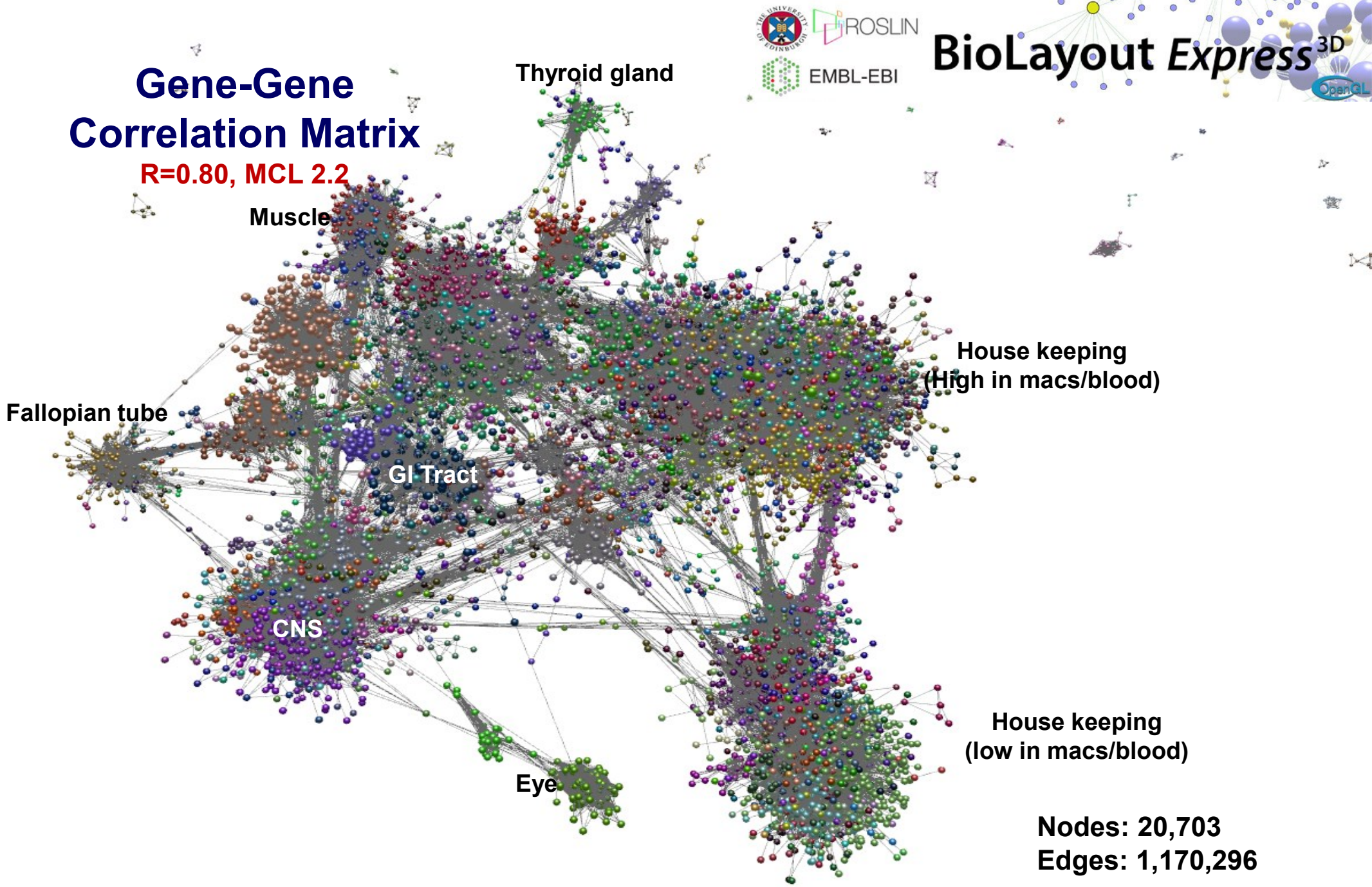
Expression profiles



http://biogps.org

Gene-Gene Correlation Matrix

R=0.80, MCL 2.2



Nodes: 20,703
Edges: 1,170,296

Those involved:

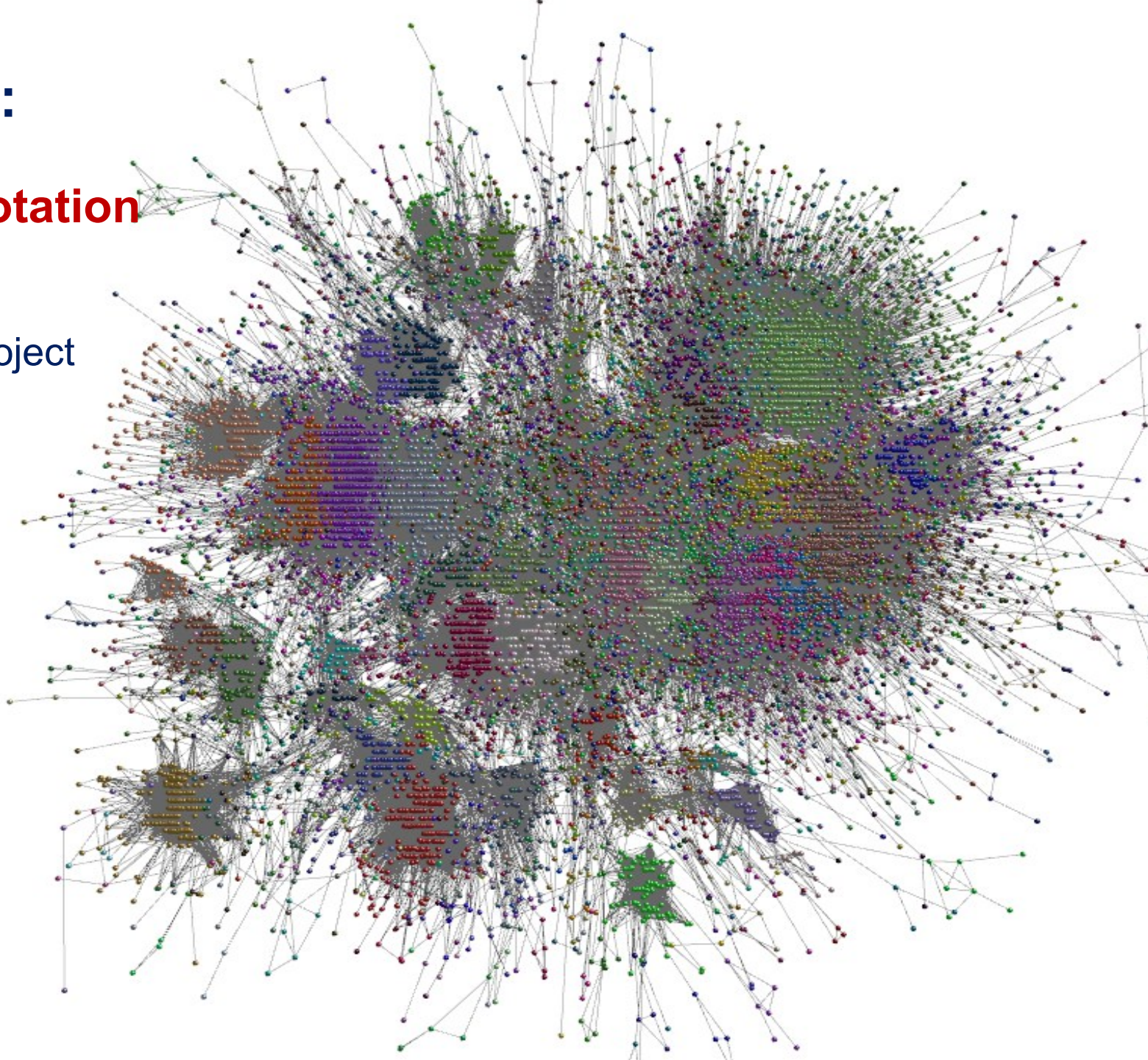
Tom Freeman

Array design/annotation

Fios Genomics
Chris Tuggle
Sanger Pig genome project
Affymetrix Inc.
Dario Beraldi

Tissue Atlas

David Hume
Alan Archibald
Mark Barnett
Kenny Bailie
Ronan Kapetanovic
Kim Summers
Lynsey Fairbairn
Andru Tomoiu
ARK-Genomics



Roslin's contribution

- Leadership
  - PiGMaP, ChickMap, ….SABRE, Quantomics, 3SR,..
  - SGSC, ISGC,…
- Genome sequence data
  - De novo: sheep (70x)
  - Re-sequence: pig, cattle, chicken
- Analysis, annotation (Ensembl)
  - Chicken, cattle, turkey,..pig,…
- SNP chip development
  - Pig, cattle, chicken, salmon,
- Microarray (ARK-Genomics)
  - cDNA to Affymetrix Snowball

ARK-GENOMICS

ARK-Genomics
Centre for Comparative & Functional Genomics

ROSLIN

- ARK-Genomics
  - High-throughput facility focusing on genetics and genomics
  - BBSRC National Capability based at The Roslin Institute
  - Collaborators on every continent

  - Offering research, collaborations and service provision

  - Investing in the latest genomics technologies
    - Sequencing
    - Genotyping
    - Transcriptomics
    - Comparative Genomics
    - Bioinformatics

http://www.ark-genomics.org

THE UNIVERSITY OF EDINBURGH

BBSRC

ARK-Genomics
Centre for Comparative & Functional Genomics

ROSLIN

Technologies

DNA Sequencing

Illumina Sequencing

- Up to 150bp paired
- Novel genomes
- Resequencing
- RNA-Seq
- ChIP-Seq
- Epigenetics

Illumina HiSeq 2000
Illumina GAIIx
Sanger 3730

Genotyping

Illumina - from HD to custom chips

- iScan, Inifinium
- BeadXpress, Goldengate
- BeadChip

Affymetrix

- GeneTitan, Axiom
- Process 96 arrays / run

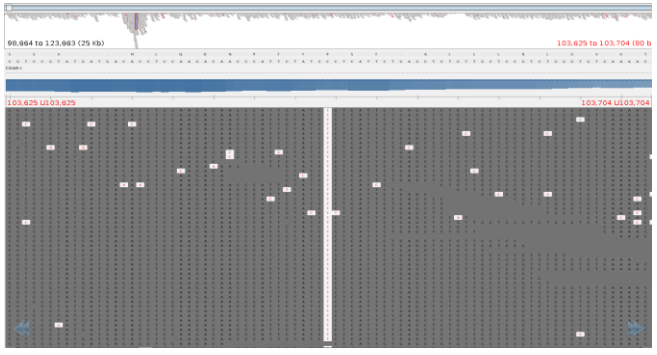
Microarrays

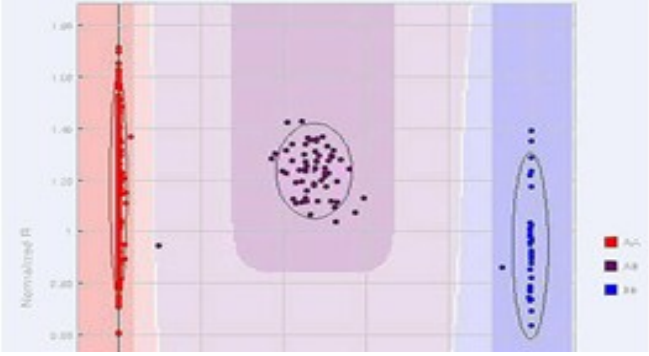
Gene Expression

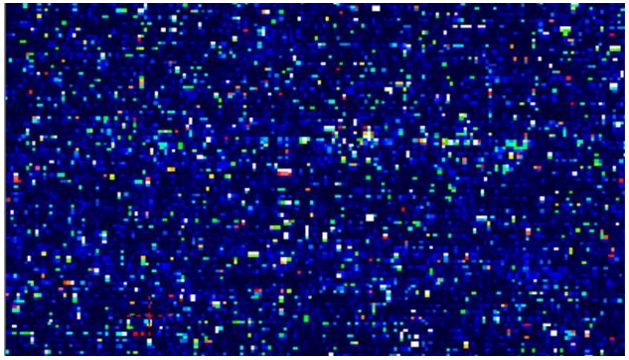
- Affymetrix
- Agilent
- Illumina
- Whole genome
- Exon-level
- microRNA

CGH, ChIP-Chip, MeIP

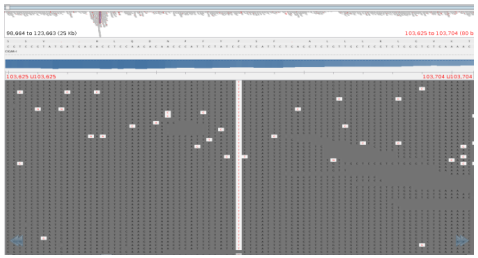
- Nimblegen

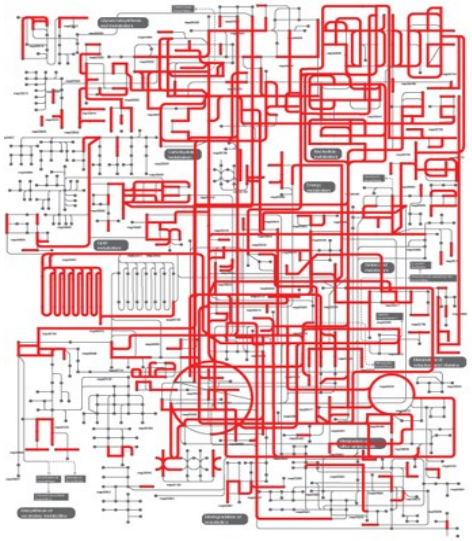


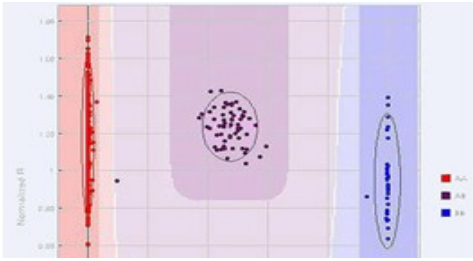




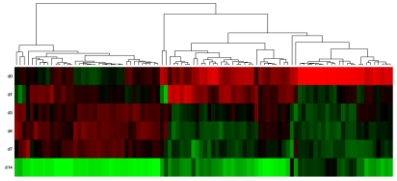
Examples of Current research







- Genome (re)sequencing
  - Sheep genome with BGI
  - 21 chicken genomes
    - With IAH: disease resistance
    - With breeders: design SNP chip
  - Bacterial genomics
- Functional genomics
  - Host-pathogen interactions, gene regulation, transcription factor binding, microRNAs, epigenetics
- Metagenomics
  - Gut microbiome: : the “forgotten organ”
  - Ruminants, chickens, others



Next Generation Sequencing – The Role of New Sequence Technologies in Shaping the Future of Veterinary Science

Hosted by the RCVS Charitable Trust

RCVS charitable trust