Technologies, resources and tools for the exploitation of the sheep and goat genomes.

B. P. Dalrymple, G. Tosser-Klopp, N. Cockett, A. Archibald, W. Zhang and J. Kijas.

The plan

- The current state of the genomes
- The current resources and tools
- The relationships between the genome assembly and the tools and resources and utility
- What do the next two years hold?
- Looking further ahead

THE STATE OF THE GENOME

Goat

- ~2.66-Gb draft genome assembly of a female Yunnan black goat
- Dong et al., Nat Biotechnol. 2013 Feb;31(2):135-41
- Super scaffolding using optical mapping

	Contig		Scaffold		Scaffold with fosmid sequences		Super-scaffold	
	bp	Number	bp	Number	bp	Number	bp	Number
N90	4,410	141,869	440,999	1,348	582,523	976	3,825,368	167
N80	7,994	100,335	846,998	922	1,175,001	664	6,481,573	119
N70	11,323	73,948	1,253,003	664	1,739,998	481	9,700,927	88
N60	14,862	54,526	1,694,371	482	2,447,724	352	13,235,657	66
N50[a]	18,720	39,408	2,212,139	344	3,057,189	254	16,328,867	49
Total[b]	**2,522,851,955**	**542,145**	**2,662,658,003**	**285,383**	**2,662,728,047**	**284,683**	**2,446,439,202**[c]	**315**

Sheep Oar v3.1

- Texel ewe (BGI), ram (Roslin, BCM)
  - Some additional data generation and data analysis from Oar v 2.0
- Primary objective to refine draft assembly of the sheep reference genome
  - Halved number of gaps - doubled contig N50
  - Scaffold order and orientation refined on the basis of sheep BACs and linkage and RH maps
  - Removed 12,000 false duplicates with length 28 Mb
  - covered ~99% of the unique genome
- Released September 2012 GenBank
  - Annotated by NCBI, Ensembl (coming very soon), UCSC (?)

The Texel female was 6 months old (provided by Jacob B. Hansen, University of Copenhagen) sequenced BGI DNA: Liver RNA: 7 tissues

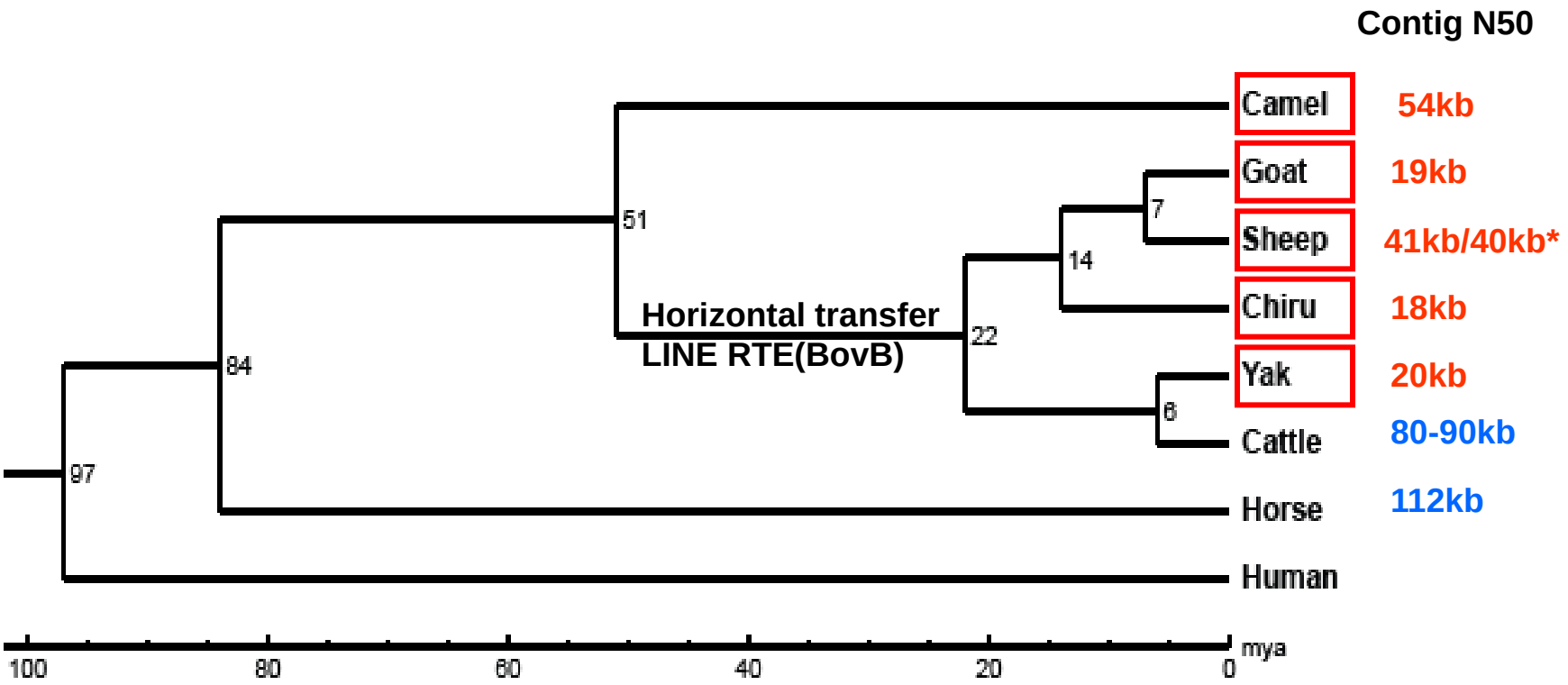
The Texel ram was used previously as the DNA source for CHORI-243 BAC library DNA: Blood (Dalrymple *et al.,* 2007).-----Roslin

Sequence used for Oar v2.0 and additional for Oar v3.1

Sample	Purpose	Sequence method	Paired-end libraries	Libraries	GA Lanes	Total length (Gb)	Reads Length (bp)	Coverage (X)
Female	assembly	Illumina	180bp	1	4	23.8	101	7.93
Female	assembly	Illumina	350bp	4	21	105.0	101	35.00
Female	assembly	Illumina	800bp	2	6	32.0	101	10.67
Female	assembly	Illumina	2kbp	2	11	35.7	45	11.90
Female	assembly	Illumina	5kb	2	6	18.5	45	6.17
Female	assembly	Illumina	9kb	1	3	8.3	45	2.77
Female	assembly	Illumina	17kb	1	1	1.8	45	0.60
Female*	fill gap	Illumina	200bp	1	1	3.0	45	1.00
Male	fill gap	Illumina	200bp	1	16	77	101	24.0
Male	fill gap	Illumina	500bp	1	24	72	101	25.5
Male**	fill gap	Illumina	554bp	8	1	36	101	12.0
Male**	fill gap	Illumina	1.3kb	1	1	27	101	9.00
Six***	fill gap	454	---			9.0	240	3.00
Male	check	454	8kb			0.7		0.60
Male	check	454	20kb			0.4		0.30
Male	Check/fill gap	Sanger	184kb			0.3	687	0.09

*MeDIP-seq for high GC content sequence
**new Illumina protocol for GC content unbiased sequence
***Six animal from breeds: AW, ROM, TEX, MER, SBF and PD

How does Oar v3.1 compare with other species?



- using next generation sequence and assembly platform
- Sanger sequence

*41 kb contigs assigned to chromosomes, 40 kb, all contigs

CURRENT PUBLIC DOMAIN TOOLS AND RESOURCES

Goats

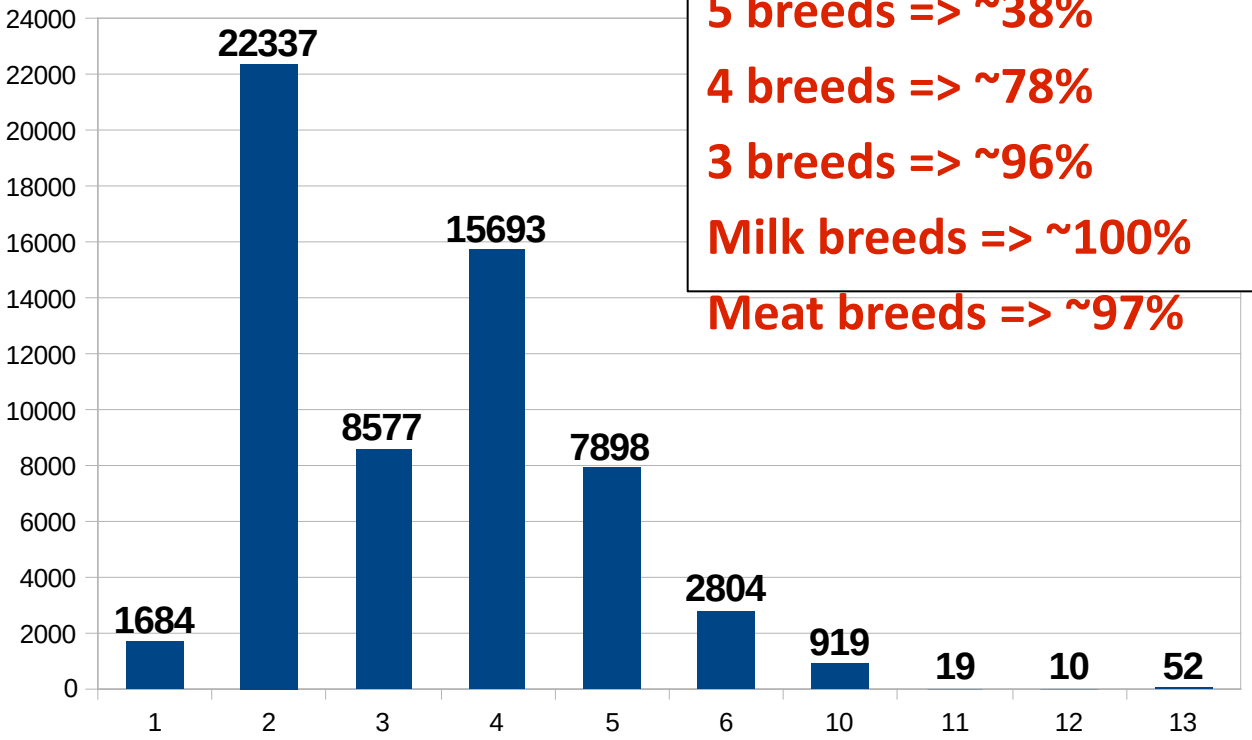
- Genome assembly
  - NCBI annotation – EST-based and predicted RefSeqs and gene models
  - Sequence searches – NCBI
  - RH panel
  - Optical Map
- 50 K SNP chip
- 25 animals resequencing
  - ADAPTmap project
    - Coordinating genotyping and sequencing goat breeds
    - http://www.goatadaptmap.org/

SNP discovery

- 6 breeds:
  - Alpine, Saanen, Creole
  - Savanna, Katjang, Boer
- Data from several laboratories
  - Genomic sequences
    - INRA, France
    - MARDI, Malaysia
    - University of Utrecht, Netherlands
  - ESTs
    - Italy, Spain, USA…

60 000 selected SNPs

1 : EST
2 : Het. in 5 breeds
3 and 4 : Het. in 4 breeds
5 and 6 : Het. in 3 breeds
10 : Het. S & A
11 : Het. (A or S)
and (C or B or KS)
12 : Het. C and (B or KS)
13 : Het. A or S
20 : other
90 : INDEL

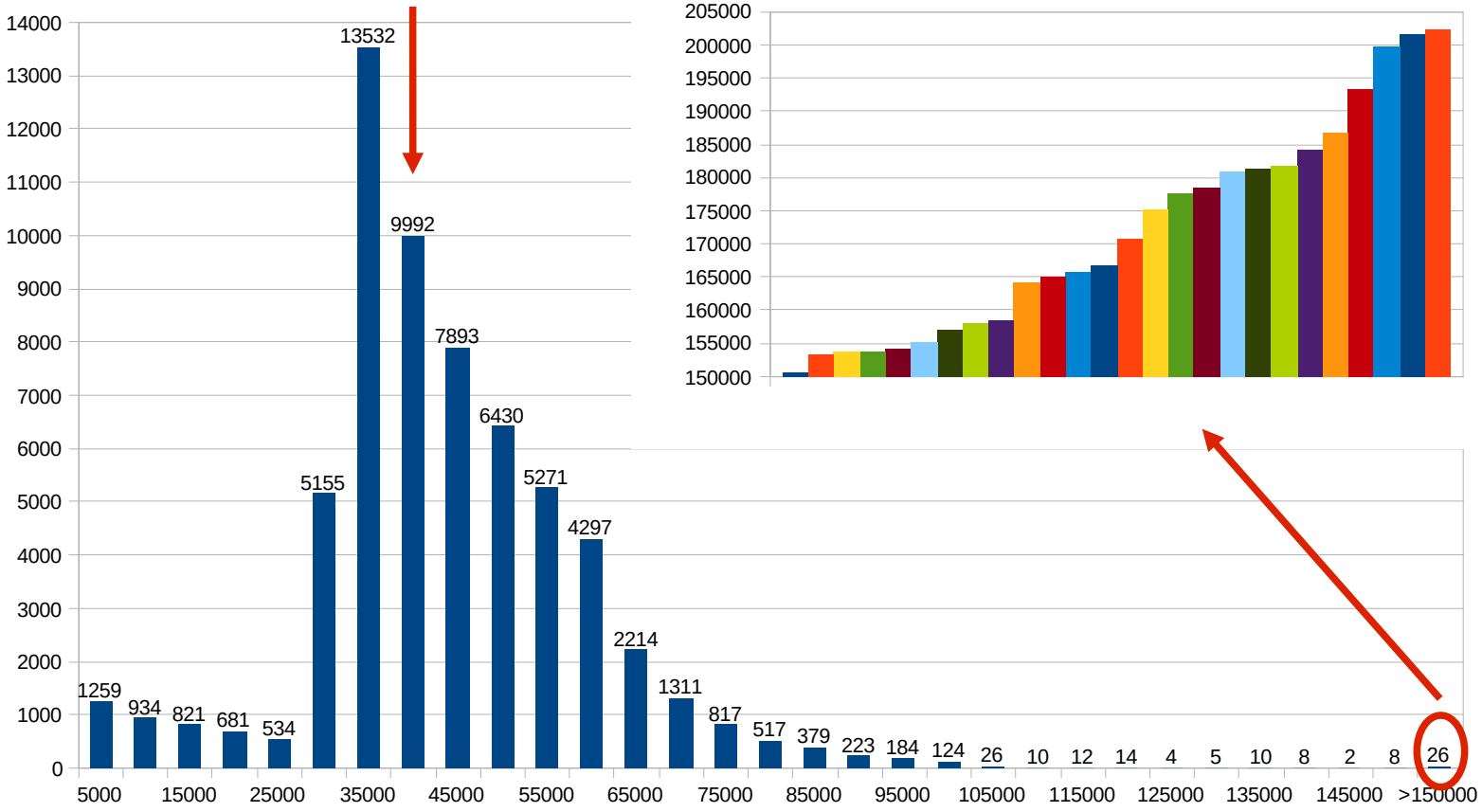


5 breeds => ~38%
4 breeds => ~78%
3 breeds => ~96%
Milk breeds => ~100%
Meat breeds => ~97%

Alpine, Saanen, Creole, Savanna, Katjang, Boer

60 000 SNPs - Spacing

- median interval => ~ 40kb



Chip manufacturing and characteristics

- Illumina iSelect design
- 288 animals were used for cluster file generation and quality control
- Includes the animals used for SNP discovery
- Breeds : Alpine, Saanen, Creole, Katjang, Savanna, Boer, Skopelos, Angora, Jinlan
- 53,348 synthesized loci
- 52,295 successful loci
- 8,000 ordered samples in September 2011
- Cluster files (.egt) available: Gwenola.Tosser@toulouse.inra.fr
- SNP sequences and annotation published in dbEST
  Information on www.goatgenome.org

A chip useful for many breeds

Breed	Samples	SNPs MAF>0.05
Alpine	53	51339
Angora	26	47195
Boer	30	48494
Creole	38	50216
Jinlan	13	45648
Katjang	13	33873
Saanen	57	51689
Savanna	20	46629
Skopelos	27	50908
Yunling	1	17335

Sheep

- Reference genome assembly
  - NCBI annotation – EST-based and predicted RefSeqs and gene models
  - Ensembl annotation – RNA-Seq-based and predicted gene models
  - Genome browsers and sequence searches – NCBI and Ensembl
  - Genes with allelically imbalanced expression
- SNP chips/panels
  - Parentage SNP panel
  - Small SNP panel- 7k
  - Ovine SNP 50K beadchip – Illumina
  - Sheep HD beadchip – Illumina
- Hapmap project SNP 50 genotypes
  - ~2800 animals diversity fo breeds, multiple animals per breed
    - Kijas et al., PLoS Biol. 2012 Feb;10(2):e1001258

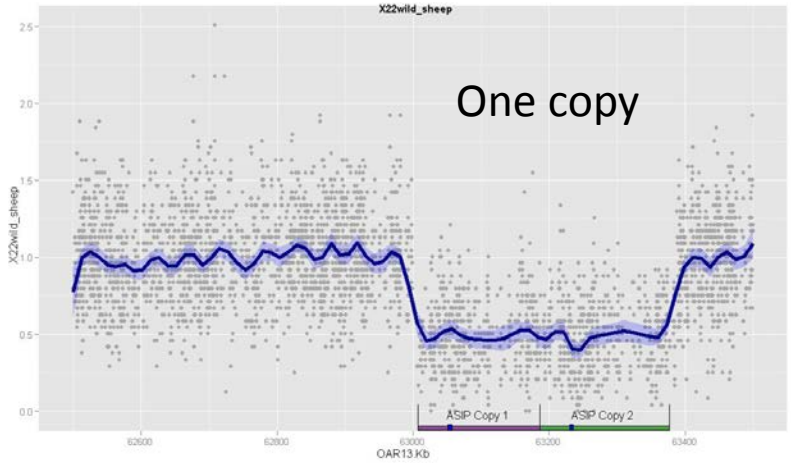
Sheep

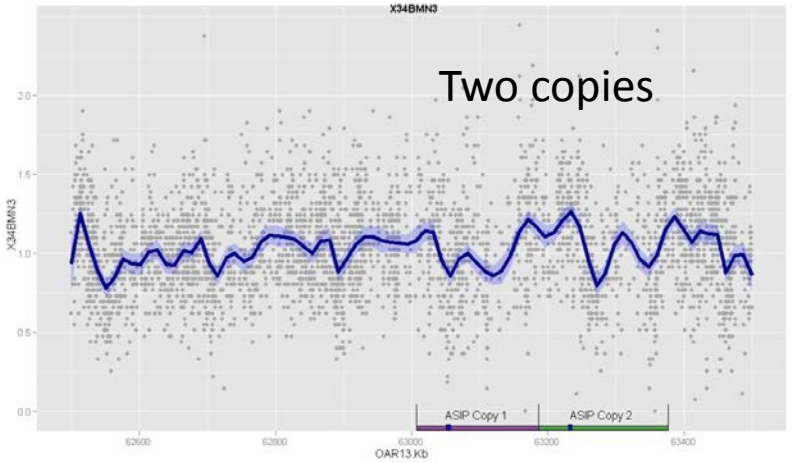
- Genome resequencing
  - >75 individuals diversity, mainly single individuals per breed
  - ~32 million SNPs
- CNV region list - preliminary
- Tissue gene expression survey ~40 tissues
  - Texel trio and fetus with same parents

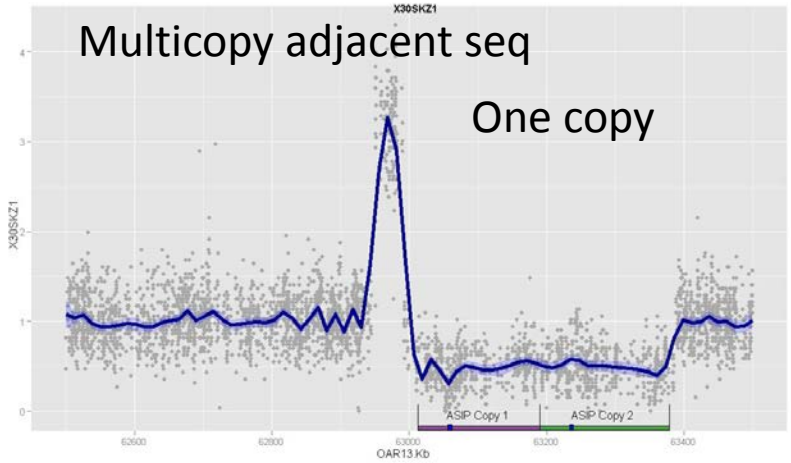
The sheep variome/resequencing

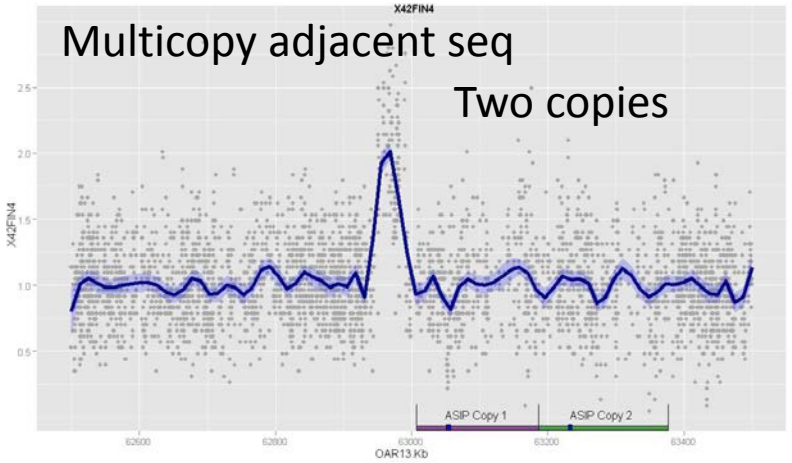
- ~10 X Illumina read coverage was generated from each of 68 diverse domestic sheep, 3 *Ovis canadensis* (bighorn) and 2 *Ovis dalli* (thinhorn) individuals.
- The reads from each animal were aligned to the reference sheep genome to identify
  - SNPs
  - Insertions
    - Reference genome does not contain all sequence present in any sheep
  - Deletions
    - All sheep do not contain all the sequence present in the reference genome
  - variable coverage between individuals and hence significant CNVs.
- Sequence, BAM and vcf files are available for all of the animals by contacting James Kijas
- Sequence data also available from GenBank

Sheep CNVs – ASIP locus, sheep colour









HD chip design

SNP type	number
Equally spaced ~5kb	569,029
Literature SNPs and Indels	289
Ovine SNP50 Beadchip SNPs	49,044
Functional SNPs maf002	30,347
Functional SNPs maf01	25,180
Genotyping by sequencing (Redrep)	21,262
Chromosome unknown	1,088
TOTAL Design	685,734
TOTAL pass	606,006
Call rate >= 0.98	589,630

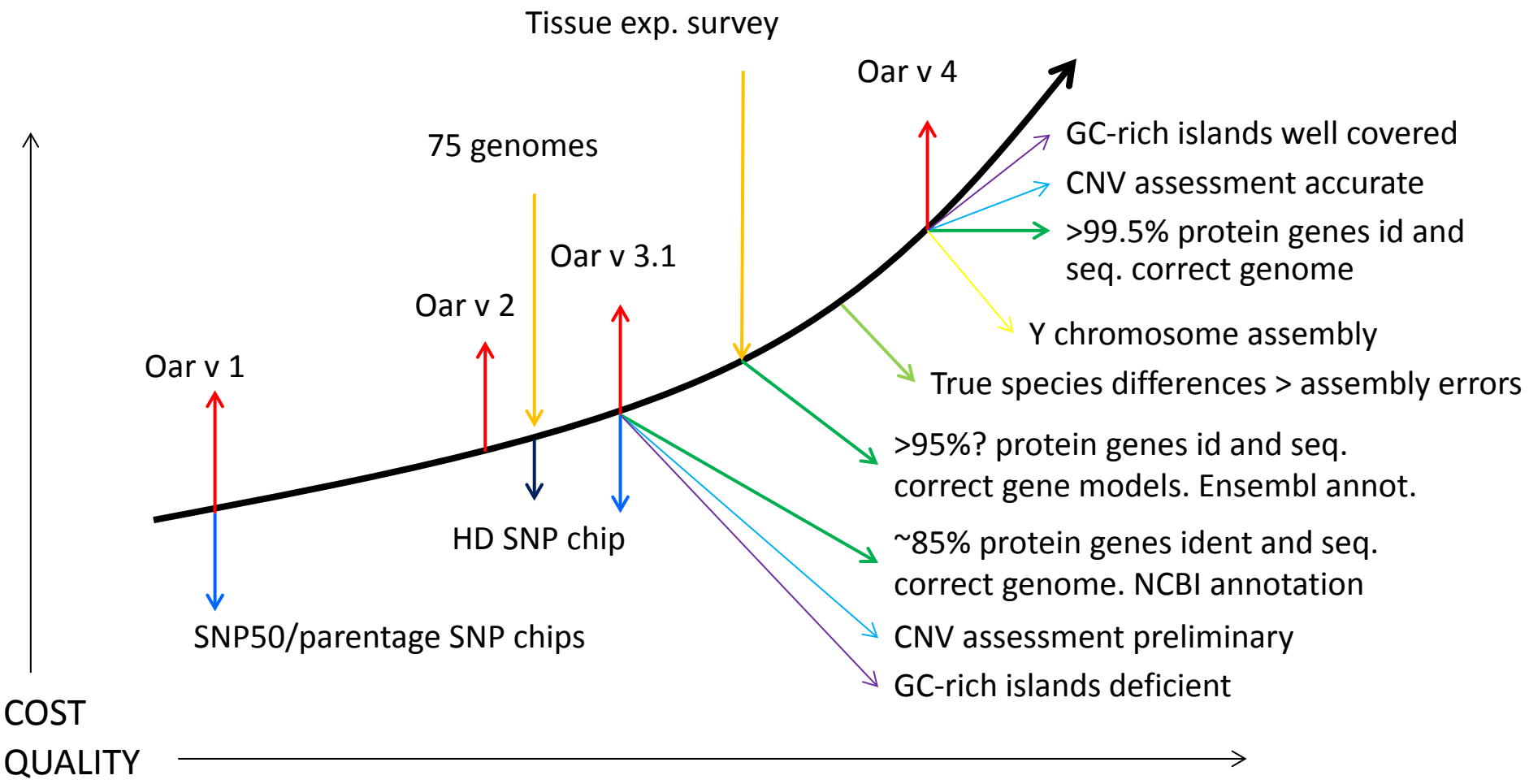
- Did not screen out SNPs in duplicated non repetitive DNA (aka CNVs and seg dups) and these can be detected (and used)
- Illumina iSelect
- Mean sample call rate = 0.9932 +/- 0.0006
- Mean SNP call rate 0.9931 +/- 0.0695, that includes all the snps with no calls

Tissue gene expression library

- 3-SR and Roslin
- Texel ram and ewe, their lamb and their embryo
- Paired end, stranded, 150 base reads
- ~40 tissues
- ~1.1 TB of data
- Forms the core of Ensembl annotation of the sheep genome

THE RELATIONSHIPS BETWEEN THE GENOME AND TOOLS

Cost v. quality v. utility



THE NEXT TWO YEARS

Goat

- genome diversity database programme / ADAPTmapADAPTmap project will allow data sharing (50K SNP chip and resequencing) for genetic diversity studies, detection of selective sweeps, detection of CNV (using 50K chips)
- Interest in fibre production - collaboration between several countries China, Australia, Argentina..
- Seeking funding for resequencing of several breeds (with 20 individuals per breed including wild breed) and detect structural variations (mate-pair libraries)
- exon database for goat/RNAseq data (28 tissues of Cashmere goats and Angora)

Sheep

- Small number of localised reassembly patches
  - From BAC sequencing
  - Improving assembly prior to PacBio overlay
- Pacbio sequencing integrated with Illumina-based assembly
  - Oar v 4
    - Current gaps halved
      - Although low coverage should fill the unique sequence gaps as these are mainly short
      - Remaining gaps likely to be the big ones, especially if repeat rich in high copy number repeats
    - Most sequencing and assembly errors corrected
    - Very high quality protein coding gene annotations
    - Suitable for methylation studies
- Sheep genomes database
  - Many more genomes/exomes sequenced

Pacbio sequencing

- Expecting better than
  - mean read length of 2.3 kb
  - half of the data is in reads longer than 4.2 kb
  - ten percent of the reads are longer than 6.2 kb
- Long reads definitely close gaps
- The better the base assembly the better the performance of PacBio in filling gaps

	D. pseudoobscura Fly	*S. purpuratus* Sea urchin	*M. undulates* Bird	*R. norvegicus* Rat	*C. atys* Monkey (preliminary)	*C. atys* Monkey (complete)
Genome Size	150 Mb	0.8 Gb	1.2 Gb	2.8 Gb	2.8 Gb	2.8 Gb
PacBio Coverage	24x	10.6x	5.1x	9.3x	6.8x	12.3x
Closed Gaps	68%	38%	21%	35%	64%	55%
Starting Gaps	6,010	142,764	49,376	109,263	186,841	115,698
Remaining Gaps	1,930	87,844	39,204	72,623	66,211	52,146
Contig N50 before	53 kb	14 kb	134 kb	60 kb	35 kb	57 kb
Contig N50 after	204 kb	24 kb	233 kb	118 kb	128 kb	166 kb

THE LONGER TERM

- Comparative genomics within and between species becomes a major force
  - Within a species, requires high quality reference genome and understanding of mechanisms of variation (CNVs etc.)
    - Pangenome of sheep
  - Between species requires true differences >> assembly problems
- Linking phenotypes to molecular mechanisms
  - Predictive models based on biological processes
- The emergence of comparative systems biology within and between species

Funding sources

- International Science Linkages – Australian Government
- New Zealand Government
- Meat and Livestock Australia
- Australian Wool Innovations
- USDA
- BGI
- Genesis Faraday
- 3-SR
- UNCEIA
- Capgenes
- Apis-gene
- FarmIQ