Mapping by RH sequencing: organizing NGS scaffolds into chromosomes

Man Rao, Mireille Morisson, Valerie Fillon, Sophie S. Leroux, Emeline Lhuillier, Diane D. Esquerre, Olivier O. Bouchez, Yinhua Huang, Alain Vignal, Ning Li, et al.

Mapping by RH sequencing: organizing NGS scaffolds into chromosomes

Man Rao[1], Alain Vignal[1], Mireille Morisson[1], Valérie Fillon[1], Sophie Leroux[1], Émeline Lhuillier[1], Diane Esquerré[1], Olivier Bouchez[1], Yinhua Huang[2], Ning Li[2] and Thomas Faraut[1]

Laboratoire de génétique cellulaire, INRA, Toulouse
State key laboratory for agro-biotechnology, China Agricultural University

January 13th 2013

Foreword

- First genome to be completely sequenced using the NGS technology
- with a scaffold N50 of 1.3Mb and 81,466 unanchored scaffolds

Whole genome sequences

species	strategy	year	depth	Scaffold N50 (Mb)	anchoring	assembly
Chicken	Hybrid	2004	6.6X	7	Genet, BAC FPC	84 ultracontigs
Dog	WGS	2005	7.5X	45	RH, cytogenetic	87 supercontigs
Horse	WGS	2007	7X	46	RH, FISH	83 scaffolds
Panda	WGS-NGS	2010	56X	1.3	not anchored	81,466 scaffolds
Turkey	WGS-NGS	2010	36X	1.5	Genetic, BAC FPC	28,261 scaffolds
Duck	WGS-NGS	2010	58X	1.2	not anchored	78,487 scaffolds

- The process of constructing complete chromosome sequences from sequence reads is a multi-step process.

Whole genome sequences

species	strategy	year	depth	Scaffold N50 (Mb)	anchoring	assembly
Chicken	Hybrid	2004	6.6X	7	Genet, BAC FPC	84 ultracontigs
Dog	WGS	2005	7.5X	45	RH, cytogenetic	87 supercontigs
Horse	WGS	2007	7X	46	RH, FISH	83 scaffolds
Panda	WGS-NGS	2010	56X	1.3	not anchored	81,466 scaffolds
Turkey	WGS-NGS	2010	36X	1.5	Genetic, BAC FPC	28,261 scaffolds
Duck	WGS-NGS	2010	58X	1.2	not anchored	78,487 scaffolds

- The process of constructing complete chromosome sequences from sequence reads is a multi-step process.
- A number of strategies have been proposed to fill the gaps between contigs and chromosomes.

Whole genome sequences

species	strategy	year	depth	Scaffold N50 (Mb)	anchoring	assembly
Chicken	Hybrid	2004	6.6X	7	Genet, BAC FPC	84 ultracontigs
Dog	WGS	2005	7.5X	45	RH, cytogenetic	87 supercontigs
Horse	WGS	2007	7X	46	RH, FISH	83 scaffolds
Panda	WGS-NGS	2010	56X	1.3	not anchored	81,466 scaffolds
Turkey	WGS-NGS	2010	36X	1.5	Genetic, BAC FPC	28,261 scaffolds
Duck	WGS-NGS	2010	58X	1.2	not anchored	78,487 scaffolds

- The process of constructing complete chromosome sequences from sequence reads is a multi-step process.
- A number of strategies have been proposed to fill the gaps between contigs and chromosomes.
- Every genome sequence needs a good map. Lewin *et al.* Genome Res (2009).

RH maps

- RH maps have been repeatedly used to construct maps and to assist genome assemblies for many species
  - Human, Mouse, Rat, Dog, Horse
  - Pig, Goat, Sheep
- We propose to use an RH mapping approach to order the duck scaffolds along the chromosomes

Duck genomic ressources

- Genetic Maps
  - 115 microsatellite markers organized in 19 linkage groups based on a Chinese resource family (Huang *et al.* 2006)
  - 91 microsatellite markers organized in 16 linkage groups (Marie-Etancelin *et al.* submitted)
- A fosmid (Moon and Magor 2004) and a BAC library (Yuan *et al.* 2006)
- An RH panel made of 90 hybrid clones (Rao *et al.* 2012)
- Genome sequences
  - Genome completely sequenced by Professor Ning Li at BGI in 2010
  - 78,487 scaffolds with and N50 of 1.2Mb

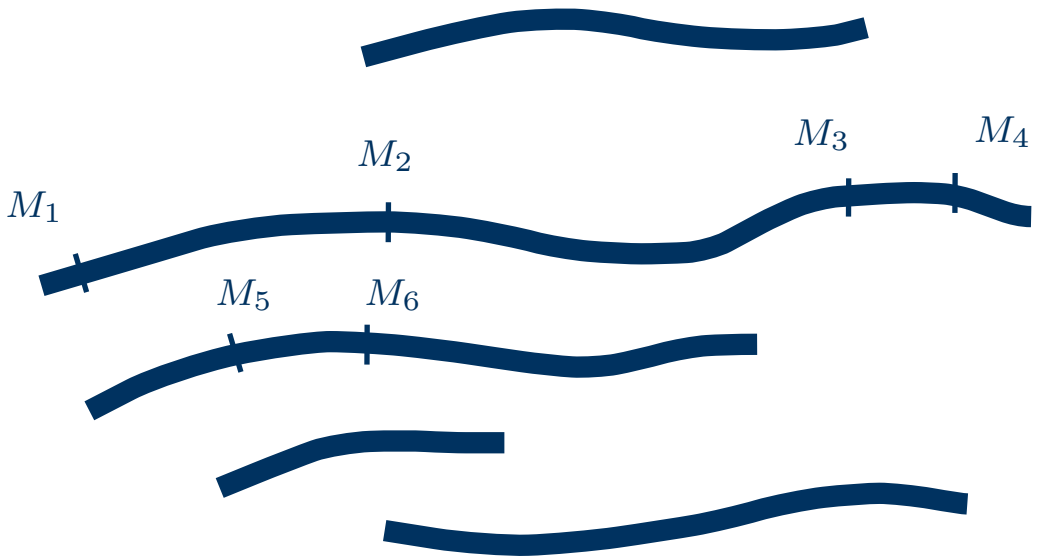
Mapping the scaffolds to chromosomes

In the absence of markers, how can we take advantage of the RH panel to order the scaffolds along the chromosomes?

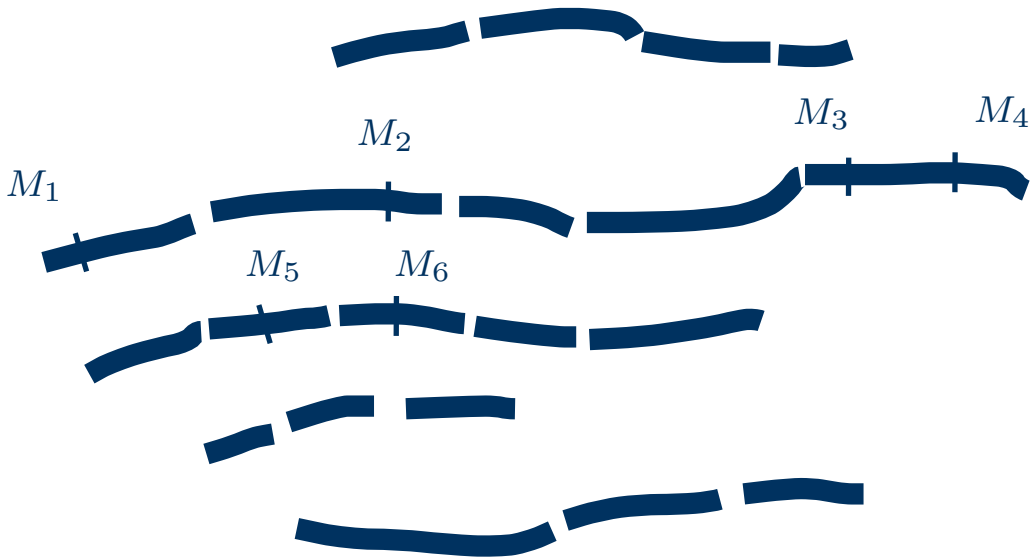
$\Rightarrow$ by sequencing the RH panel

Principle: only the retained fragments will be sequenced

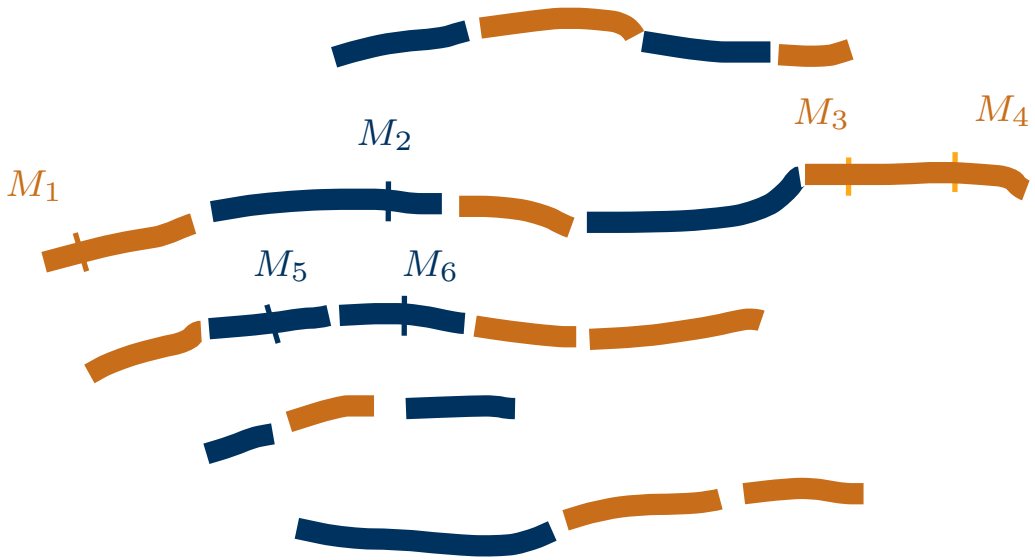
RH mapping: the principle



RH mapping: the principle



RH mapping: the principle



along the chromosomes

M_1	M_2	M_3	M_4	M_5	M_6
0	1	0	0	1	1

RH mapping: the principle

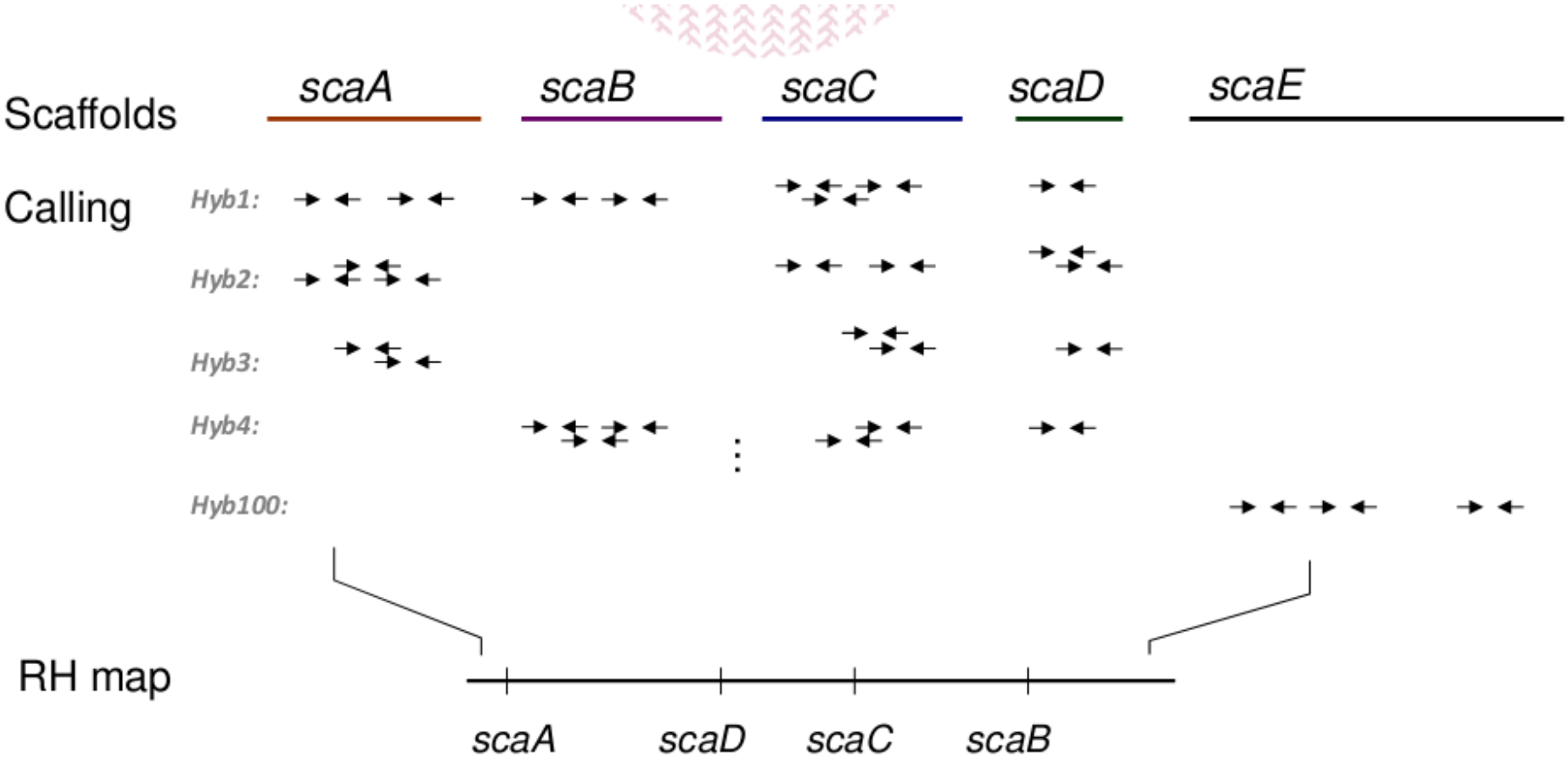
	M_1	M_2	M_3	M_4	M_5	M_6
	0	1	0	0	1	1
	1	1	0	0	0	1
	1	0	1	1	0	0
~ 100	⋮	⋮	⋮	⋮	⋮	⋮
	0	1	1	1	0	0
	0	1	0	0	1	1
	1	1	0	0	1	0

RH mapping by sequencing

We propose to sequence 96 hybrid clones of the panel using the scaffolds as markers

RH mapping by sequencing

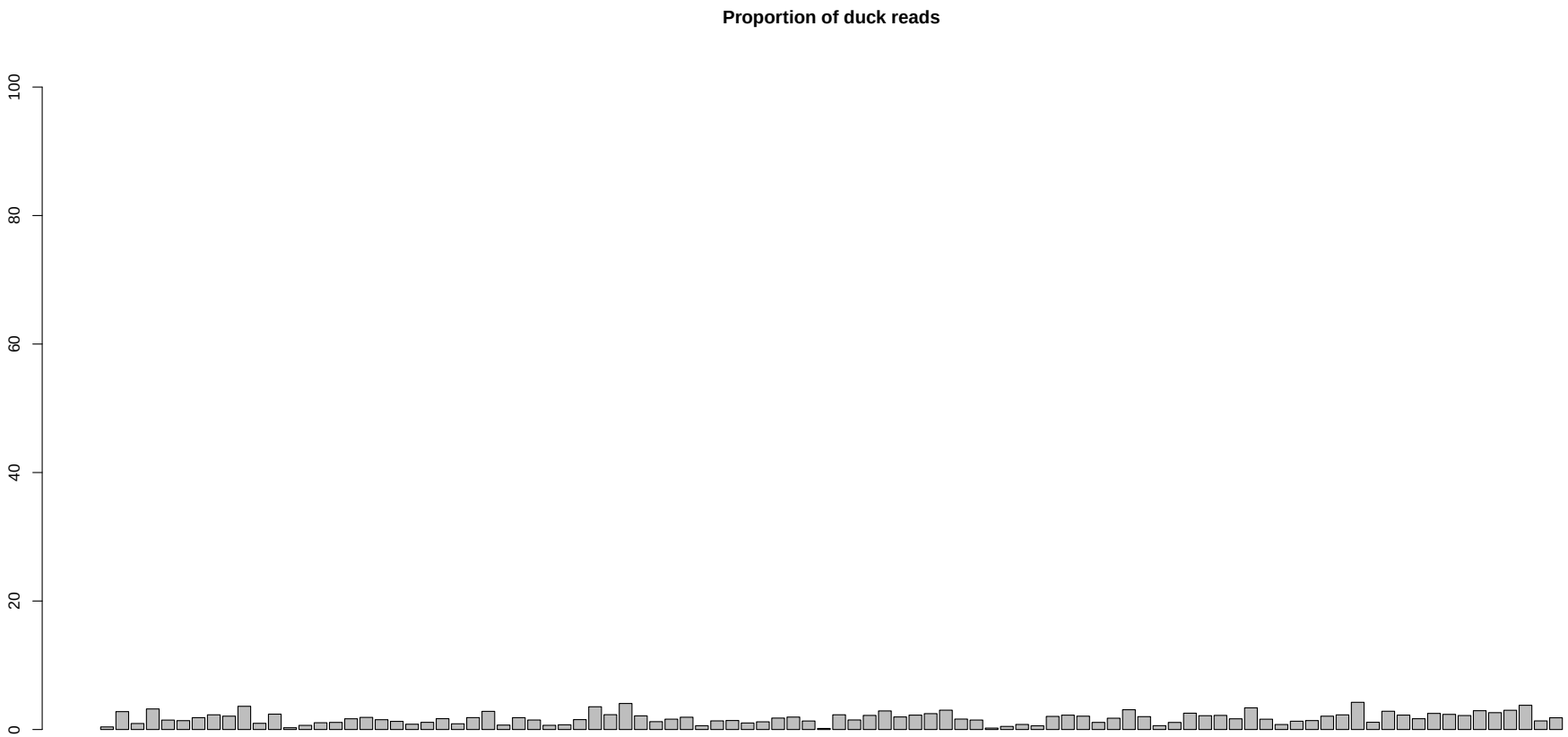
We propose to sequence 96 hybrid clones of the panel using the scaffolds as markers



Protocol

- Construction of 96 libraires, one for each hybrid clone
- Sequencing the 96 librairies, muliplexing 12 librairies per lane
- Align the reads against the Duck scaffolds
- Only the reads aligned in a proper pair are kept

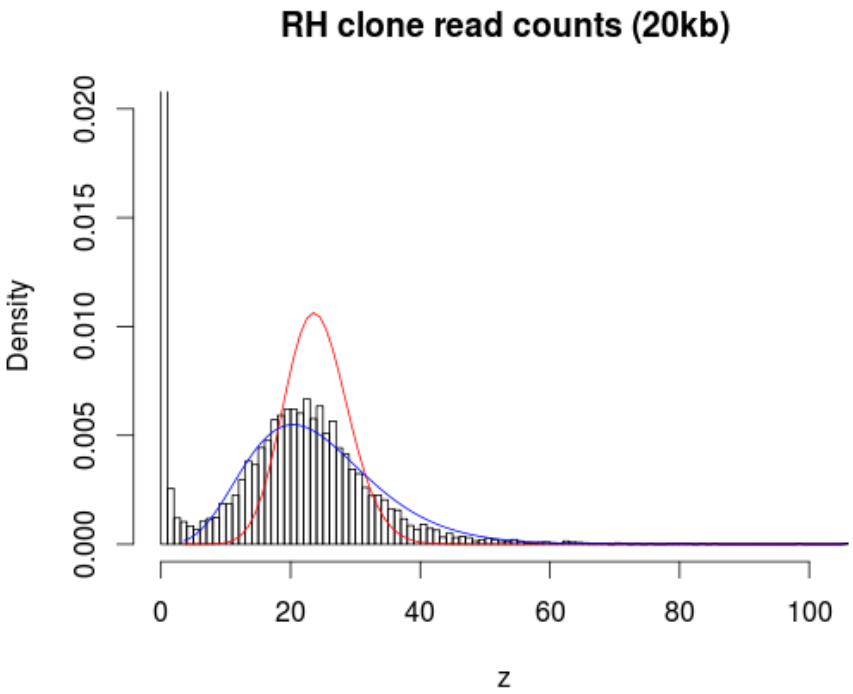
The useful reads

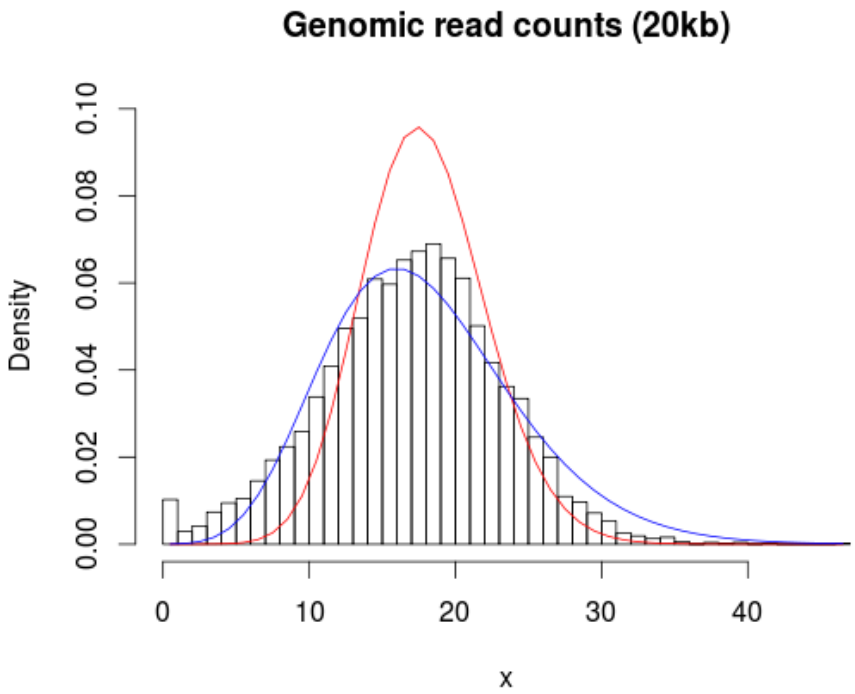


- 179 Gb produced in total
- $\sim$ 2.5% of reads can be uniquely aligned to the duck scaffolds
- $\sim$ 3% were expected: each clone contains on average 200Mb of duck genome (20% of the 1Gb duck haploid genome) and 6Gb of diploid hamster genome.
- The average sequence coverage is 0.3 X in each hybrid clone, $\sim$ 3 reads/20kb

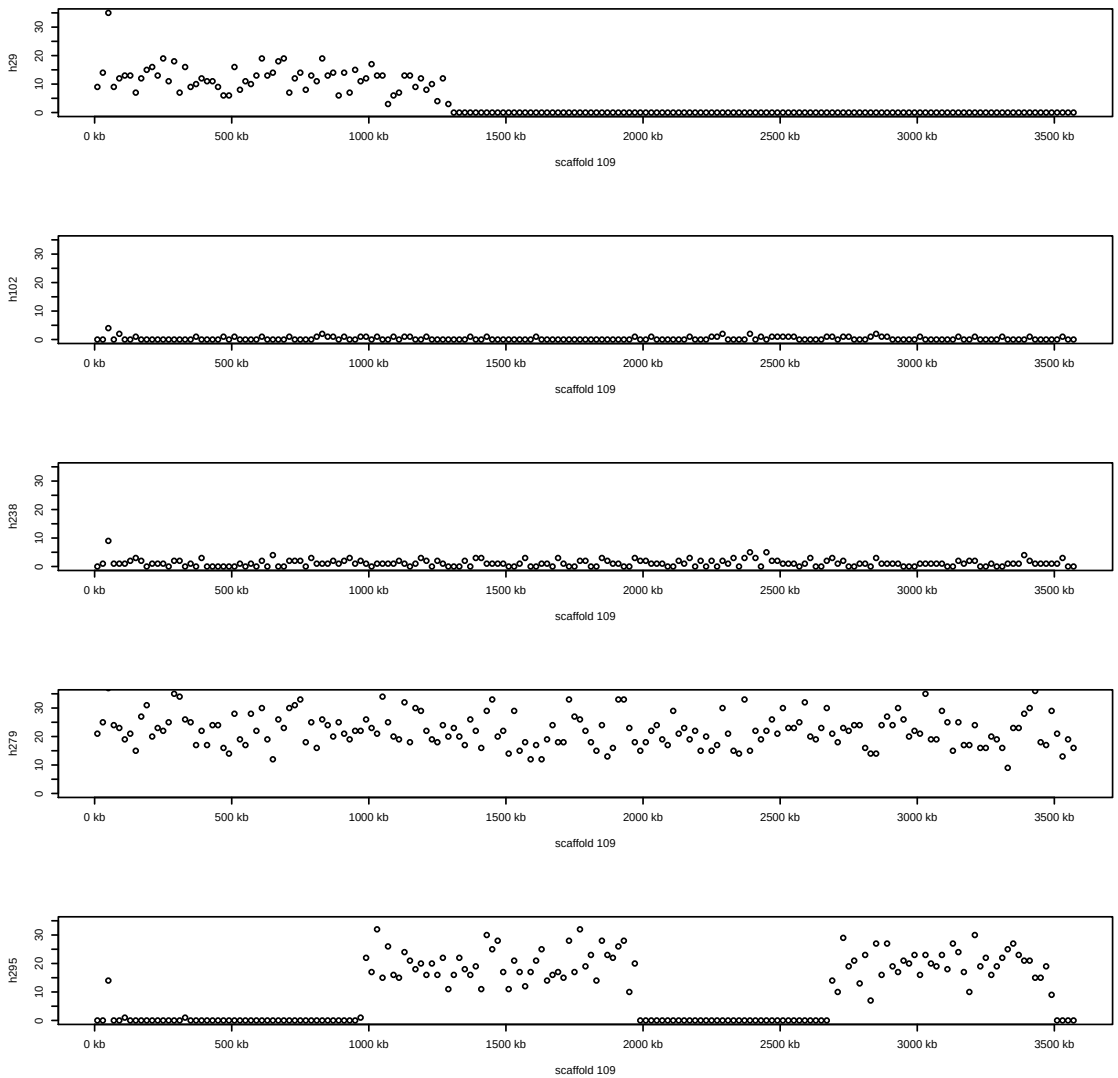
Read counts

The distribution of reads exhibits a high level of dispersion

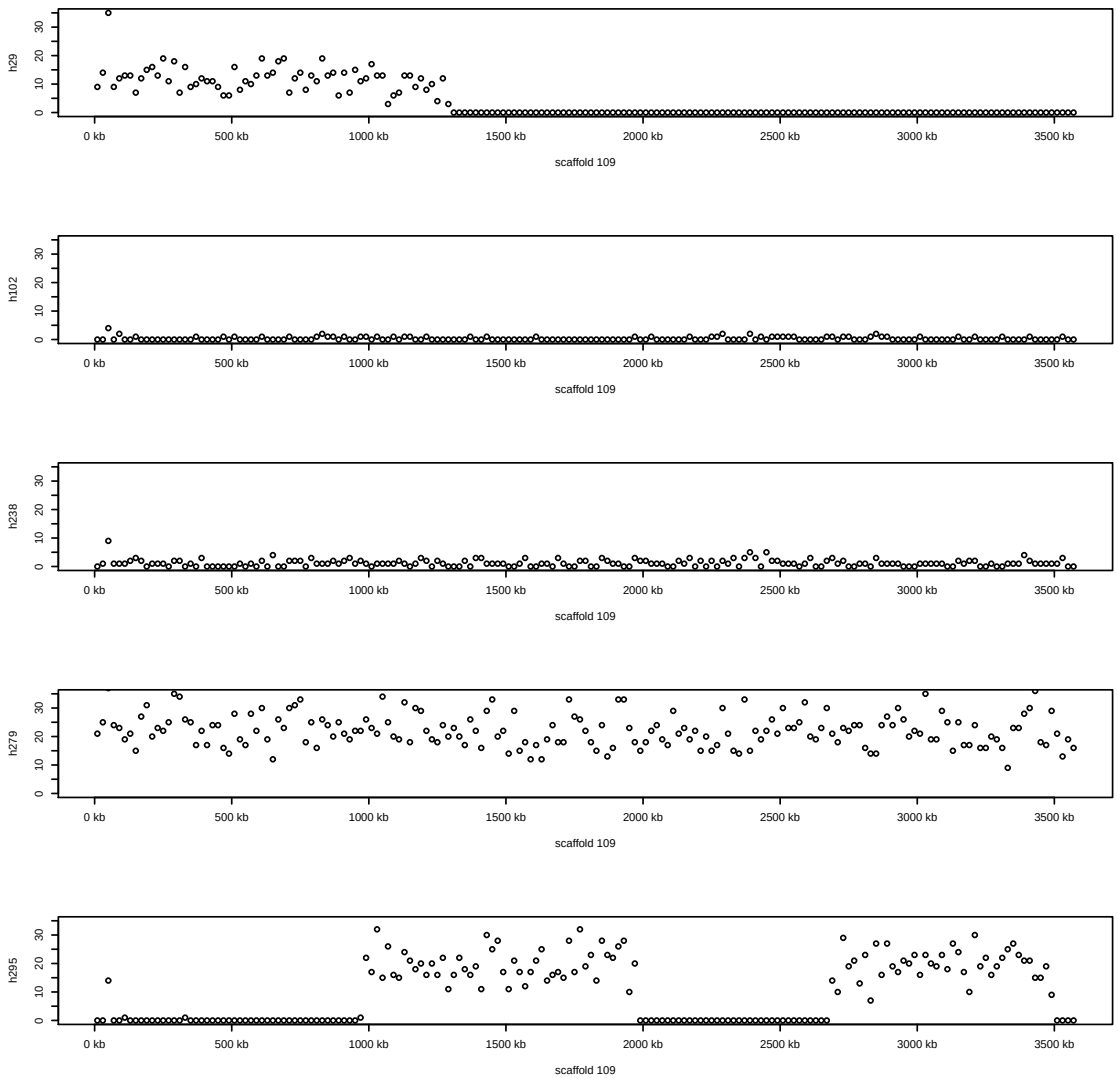




Read depth in 20kb windows for a single scaffold



Read depth in 20kb windows for a single scaffold



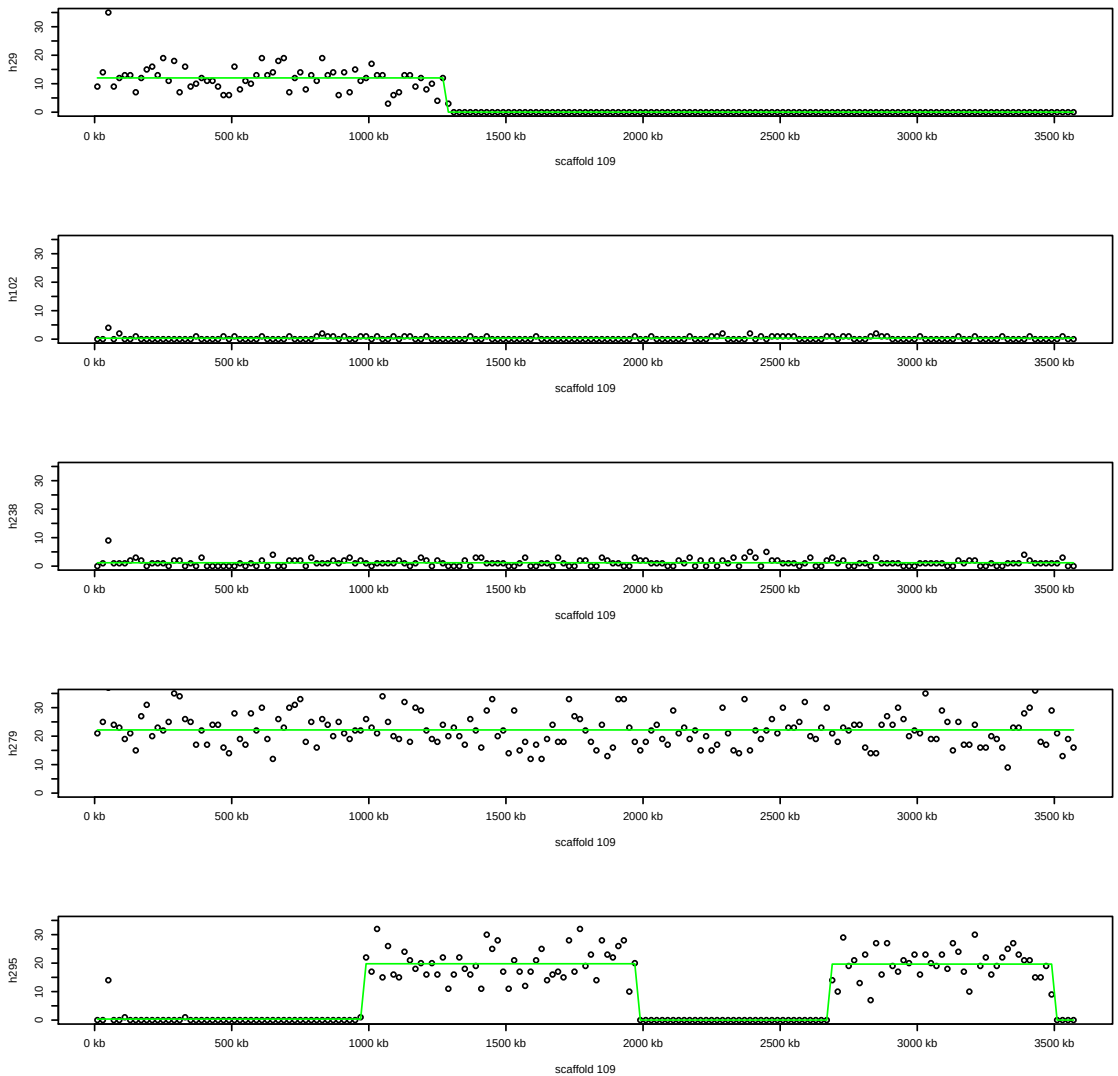
There is a need for segmentation !

Genotyping procedure

- Perform a segmentation for each scaffold in each clone using Circular Binary Segmentation (CBS) (R package DNACopy)
- Call the different segments based on the observed distribution of read counts in 20kb windows
- Construct for each scaffold a pair of markers: *left* and *right*

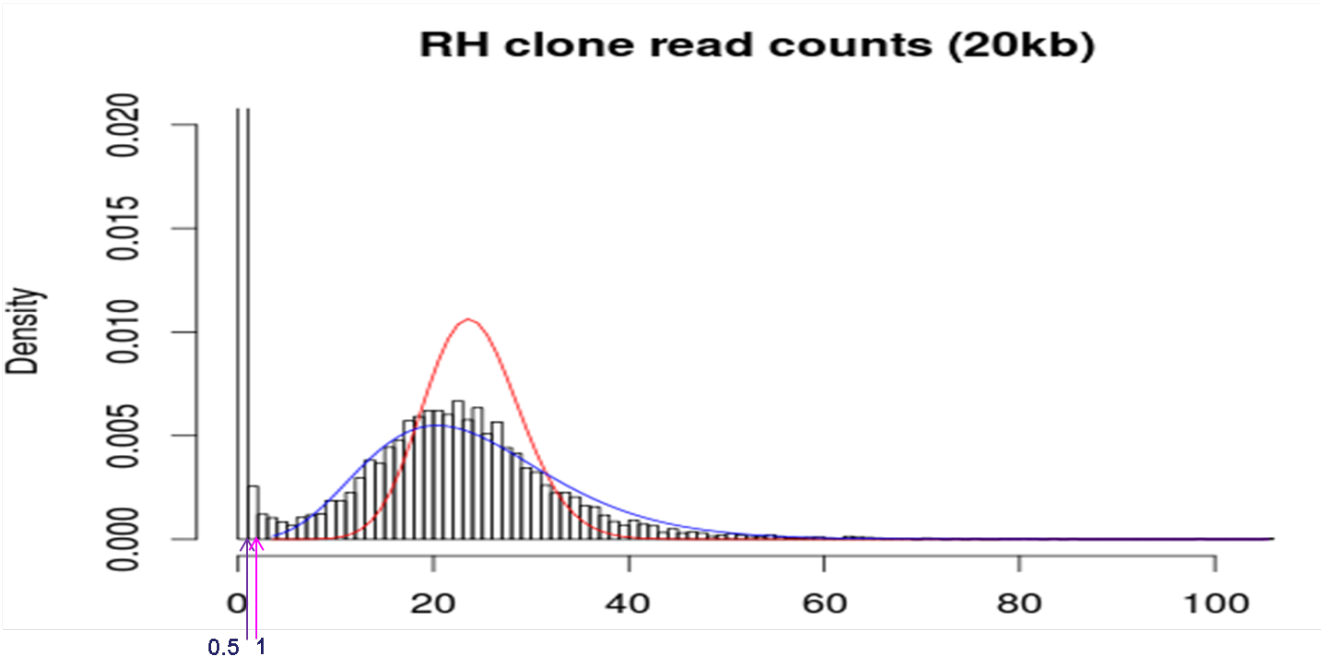
Genotyping procedure

Read depth in 20kb windows and corresponding segments for a single scaffold



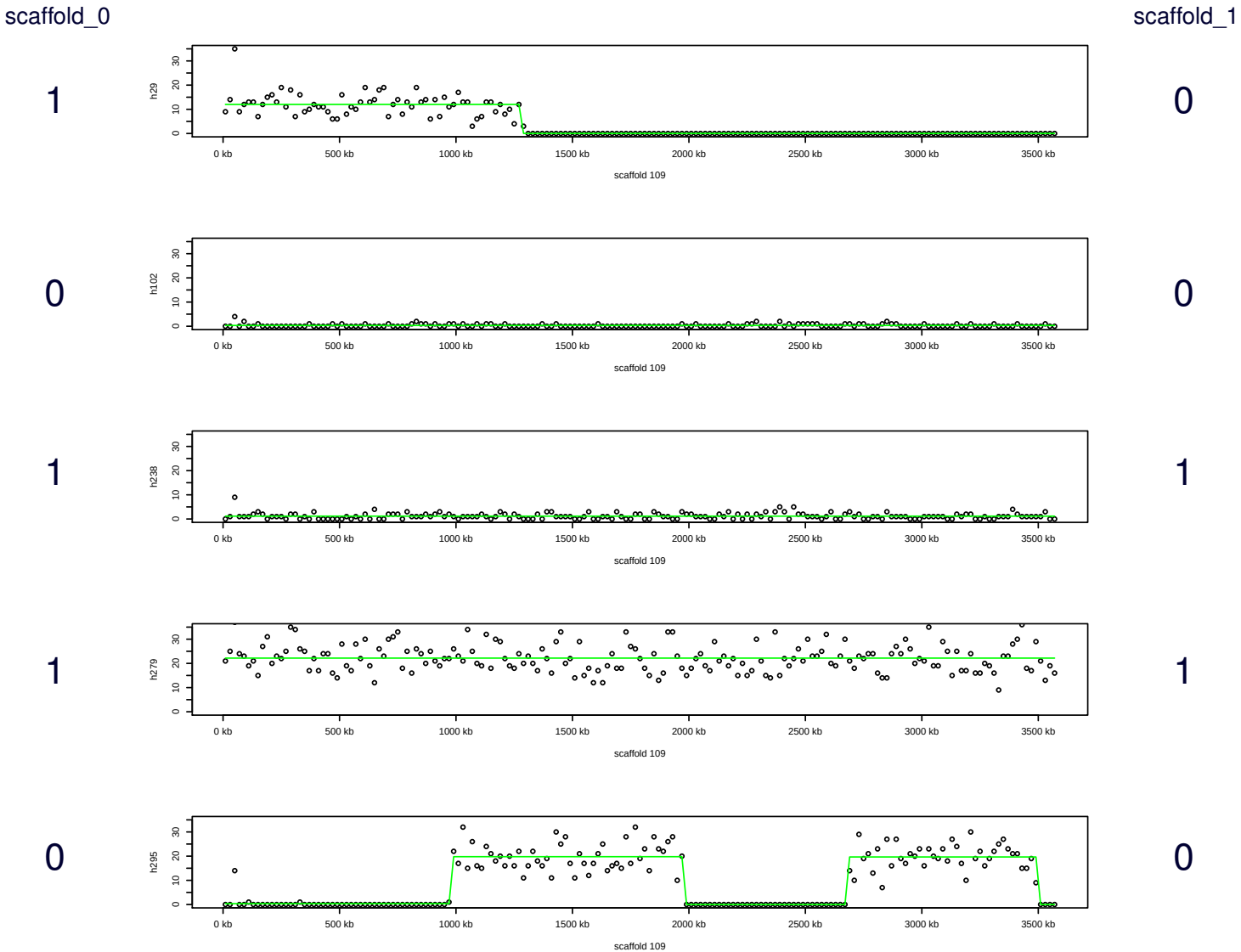
Read counts

The observed distribution of read counts is used to determine cutoff values for the presence or absence of a segment



Genotyping procedure

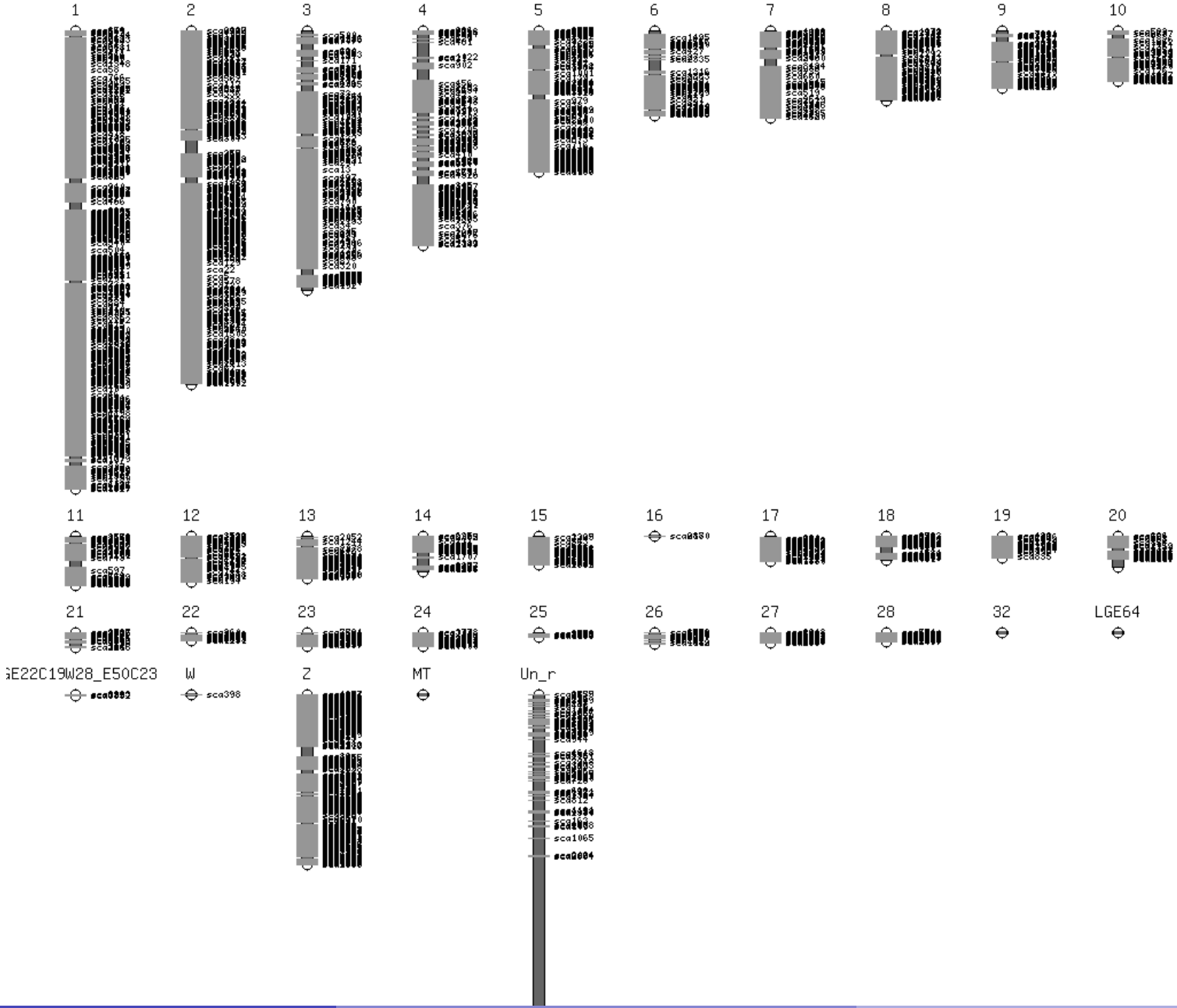
Read depth in 20kb windows and corresponding segments for a single scaffold



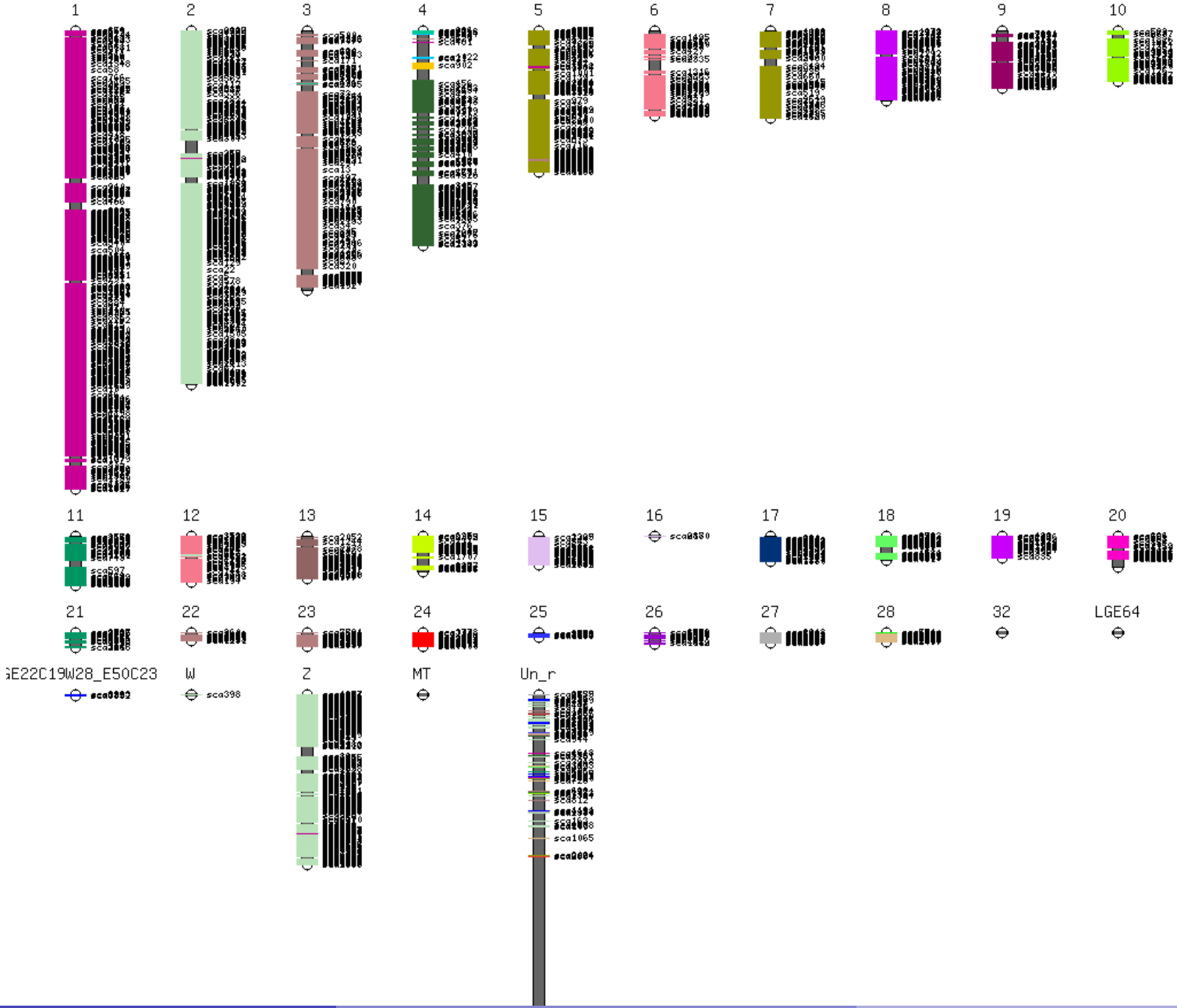
Results (1): the vectors

- RH vectors could be obtained for 2027 scaffolds totalizing 1Gb of sequences
- 675 scaffolds had a different genotype at both extremities in a least one hybrid, the rest 1352 scaffolds are treated as single marker, making a total of 2702 markers
- 1787 scaffolds could be assigned a position on the chicken chromosome

Results(2): 32 linkage groups



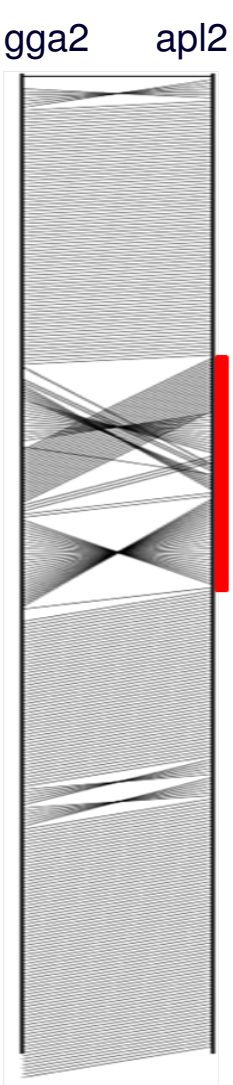
Results(2): 32 linkage groups

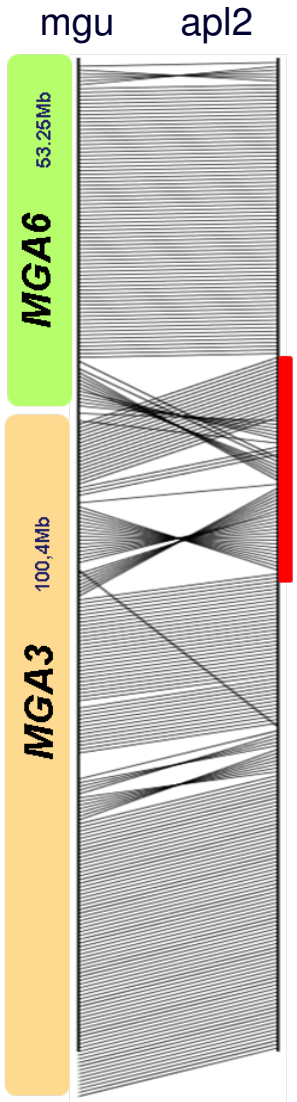


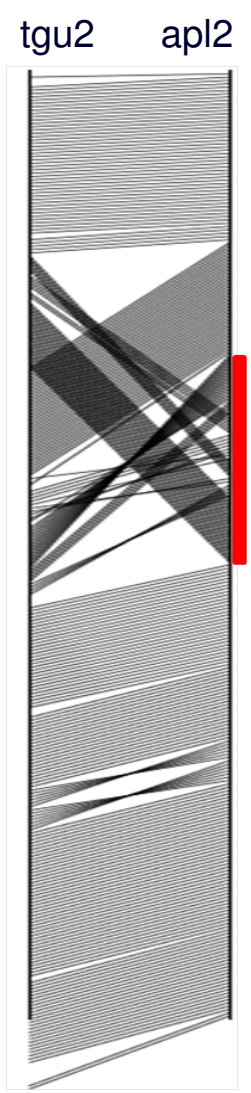
The mapping procedure

- We used a comparative approach to order the markers within each linkage group (Faraut *et al.* 2007)
- This approach also enabled us to construct robust maps (Servin *et al.* 2010)
- 2308 markers were ordered on 28 chromosomes among which 1844 are on robust maps

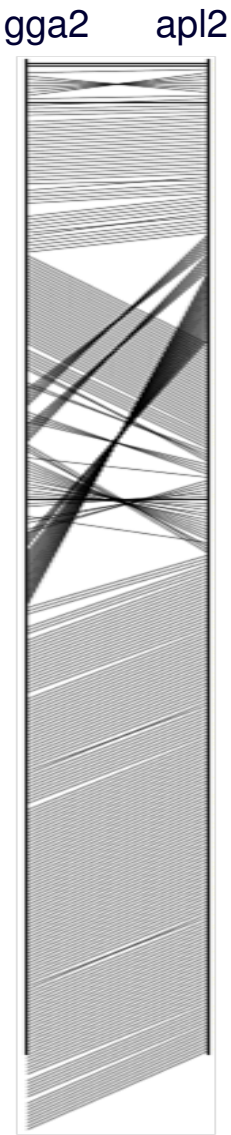
An example of duck RH map: chromosome 2

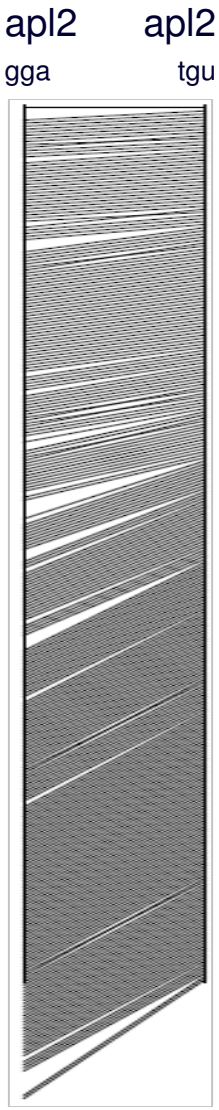






Influence of the reference genome

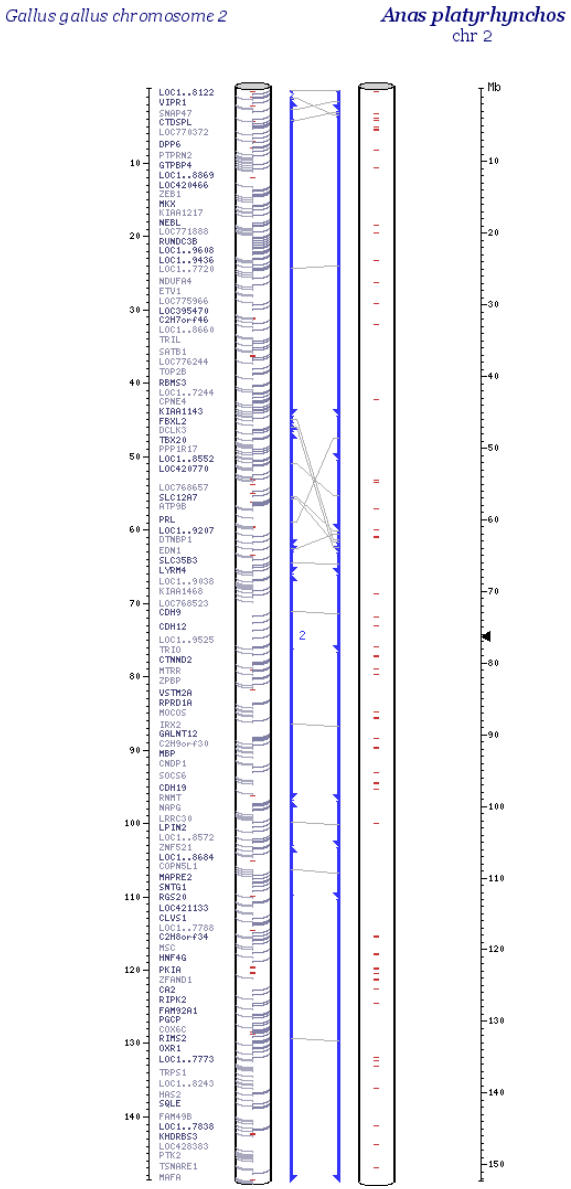




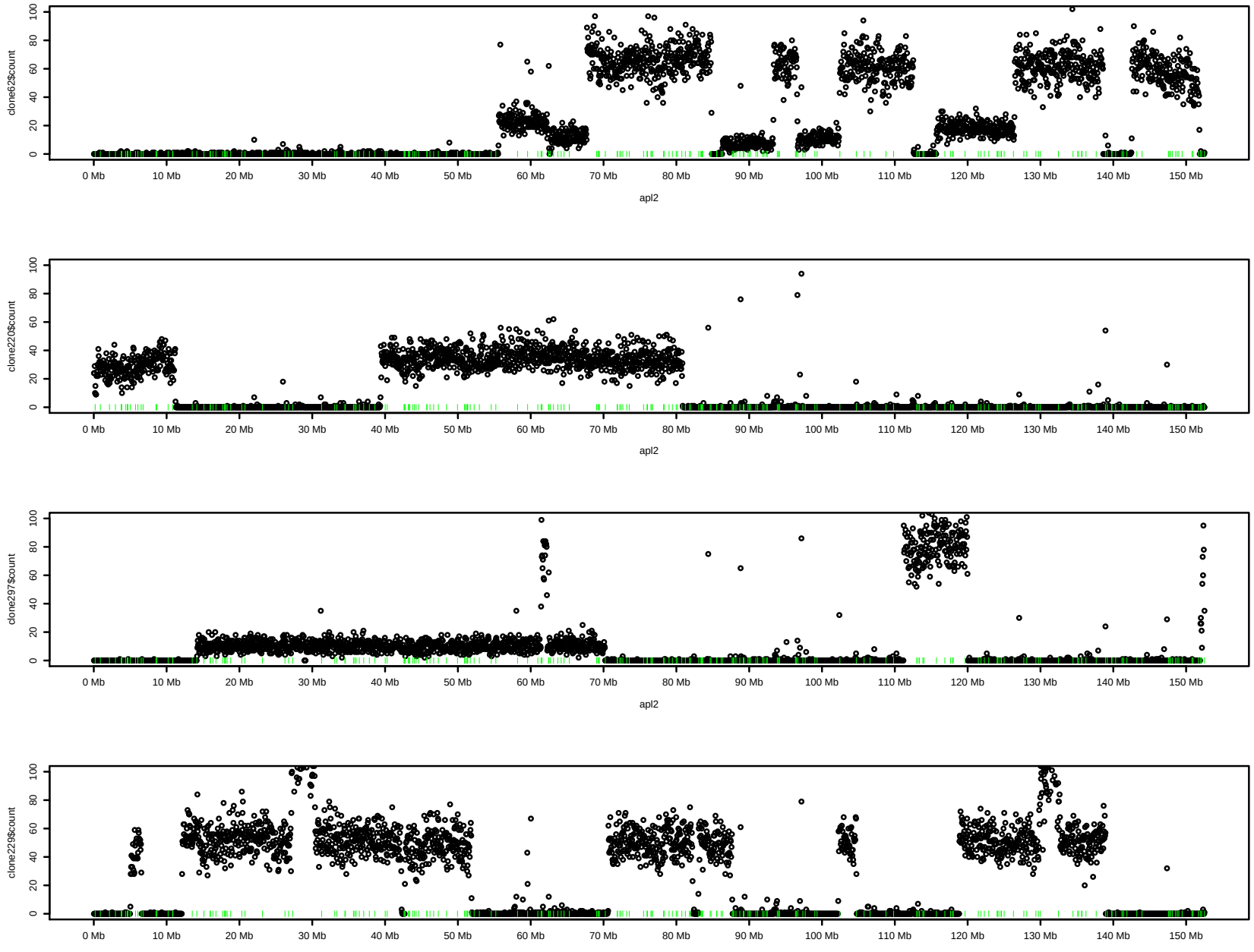
From RH maps to pseudomolecules

- The RH maps were subsequently used to construct pseudomolecules
  - Ordering the scaffolds using the maps
  - Orienting the scaffolds using the marker pairs or orientation provided by the alignment with chicken
- example of the chromosome 2: 225 scaffolds for a total of 152Mb

The apl2 pseudomolecule



Read depth on apl2 pseudomolecule



Perspectives

- Constructing pseudomolecules for each chromosome
- Mapping by Happy panel sequencing: gettint rid of the hamster
- Single sperm cell sequencing for genetic mapping

Who did what?

- Radiation hybrid panel construction: Man Rao and Mireille Morisson
- Library construction and sequencing: Sophie Leroux, Olivier Bouchez, Diane Esquerré and Emeline Lhuillier
- Genome scaffolds: kindly provided by Yinhua Huang and Ning Li
- FISH analysis: Man Rao, Valérie Fillon
- Data analysis and map construction: Man Rao, Thomas Faraut, Alain Vignal
- Speaker : Thomas Faraut instead of Man Rao

Acknowledgements

- Chinese governement for scholarship
- Animal Genetic division at INRA
- Get Plateform, Toulouse France