Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ)

Martin Mascher

IPK Gatersleben

PAG XXII

January 14, 2012

Member of the Leibniz Association

IPK GATERSLEBEN

Proof-of-principle in barley

- Diploid model for wheat
- 5 Gb genome, 80 % repetitive
- genome sequencing in progress
- physical map published last year

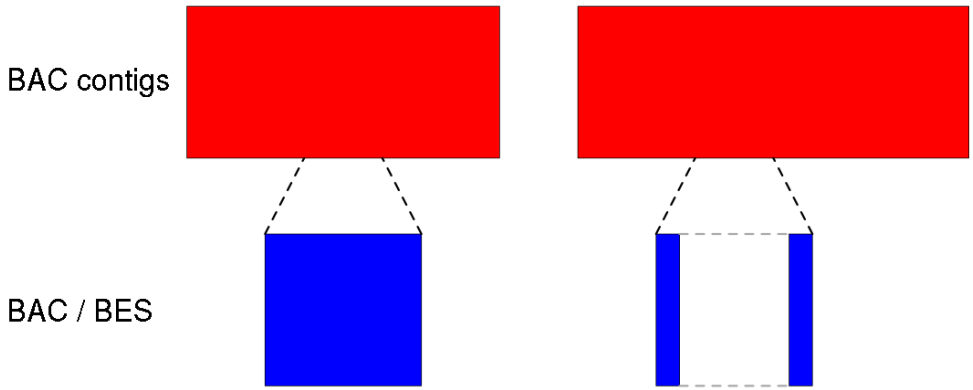
INTERNATIONAL BARLEY SEQUENCING CONSORTIUM

The sequence-enriched physical map of barley

- A physical map consisting of 9,265 BAC contigs was constructed.

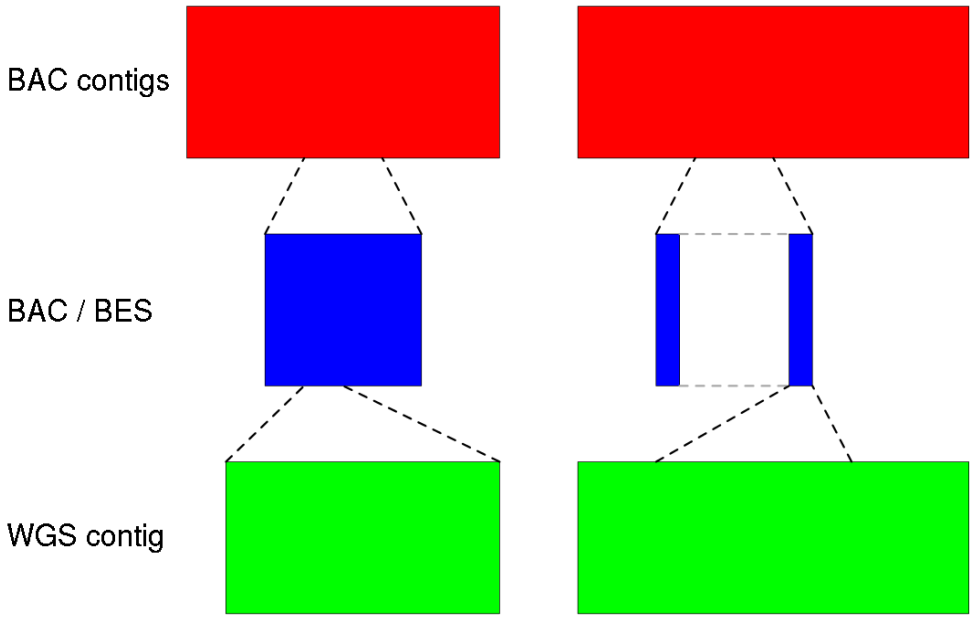
The sequence-enriched physical map of barley

- A physical map consisting of 9,265 BAC contigs was constructed.
- More than 300,000 BACs were end sequenced and $\approx$ 6,300 clones were fully sequenced.



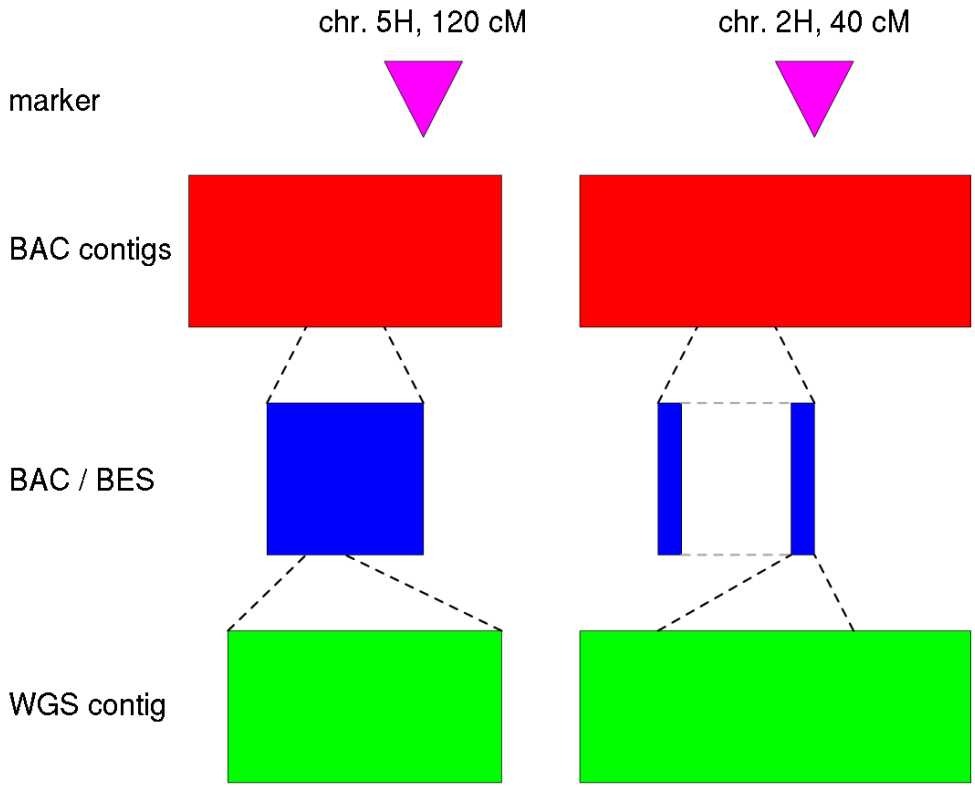
The sequence-enriched physical map of barley

- A physical map consisting of 9,265 BAC contigs was constructed.
- More than 300,000 BACs were end sequenced and ≈ 6,300 clones were fully sequenced.
- Short read data was assembled into more than 350,000 contigs larger than 1kb.



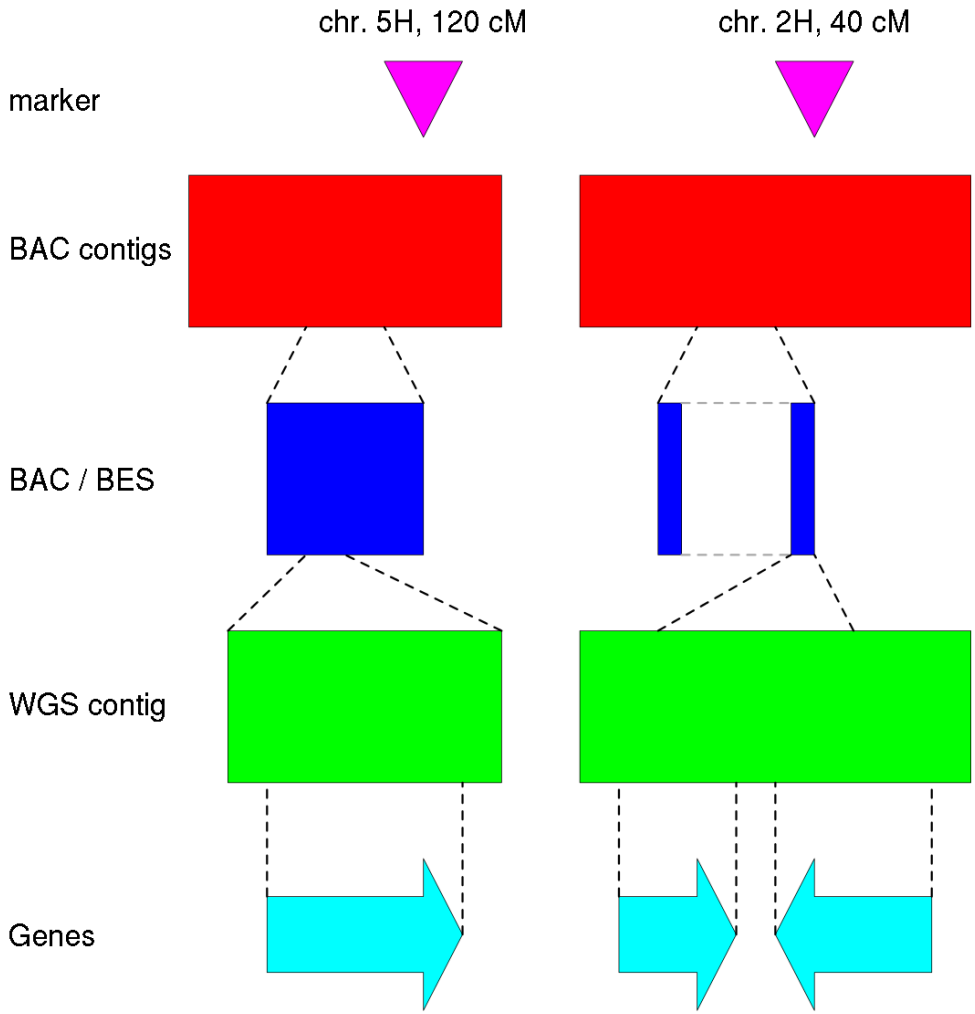
The sequence-enriched physical map of barley

- A physical map consisting of 9,265 BAC contigs was constructed.
- More than 300,000 BACs were end sequenced and ≈ 6,300 clones were fully sequenced.
- Short read data was assembled into more than 350,000 contigs larger than 1kb.
- 4,556 BAC contigs were anchored to a chromosomal location with genetic markers.



The sequence-enriched physical map of barley

- A physical map consisting of 9,265 BAC contigs was constructed.
- More than 300,000 BACs were end sequenced and ≈ 6,300 clones were fully sequenced.
- Short read data was assembled into more than 350,000 contigs larger than 1kb.
- 4,556 BAC contigs were anchored to a chromosomal location with genetic markers.



Can we do better? The idea of POPSEQ

- Only 410 Mb could be positioned in the physical framework.

no. of contigs	2.7 million
cumulative length	1.6 Gb
mean contig length	700 bp
no. contigs $>$ 1kb	376,261
lengh of contig $>$ 1kb	1.1 Gb
N50	1,425 bp

Can we do better? The idea of POPSEQ

- Only 410 Mb could be positioned in the physical framework.
- The number of genetic markers limits anchoring efficiency.

no. of contigs	2.7 million
cumulative length	1.6 Gb
mean contig length	700 bp
no. contigs $>$ 1kb	376,261
lengh of contig $>$ 1kb	1.1 Gb
N50	1,425 bp

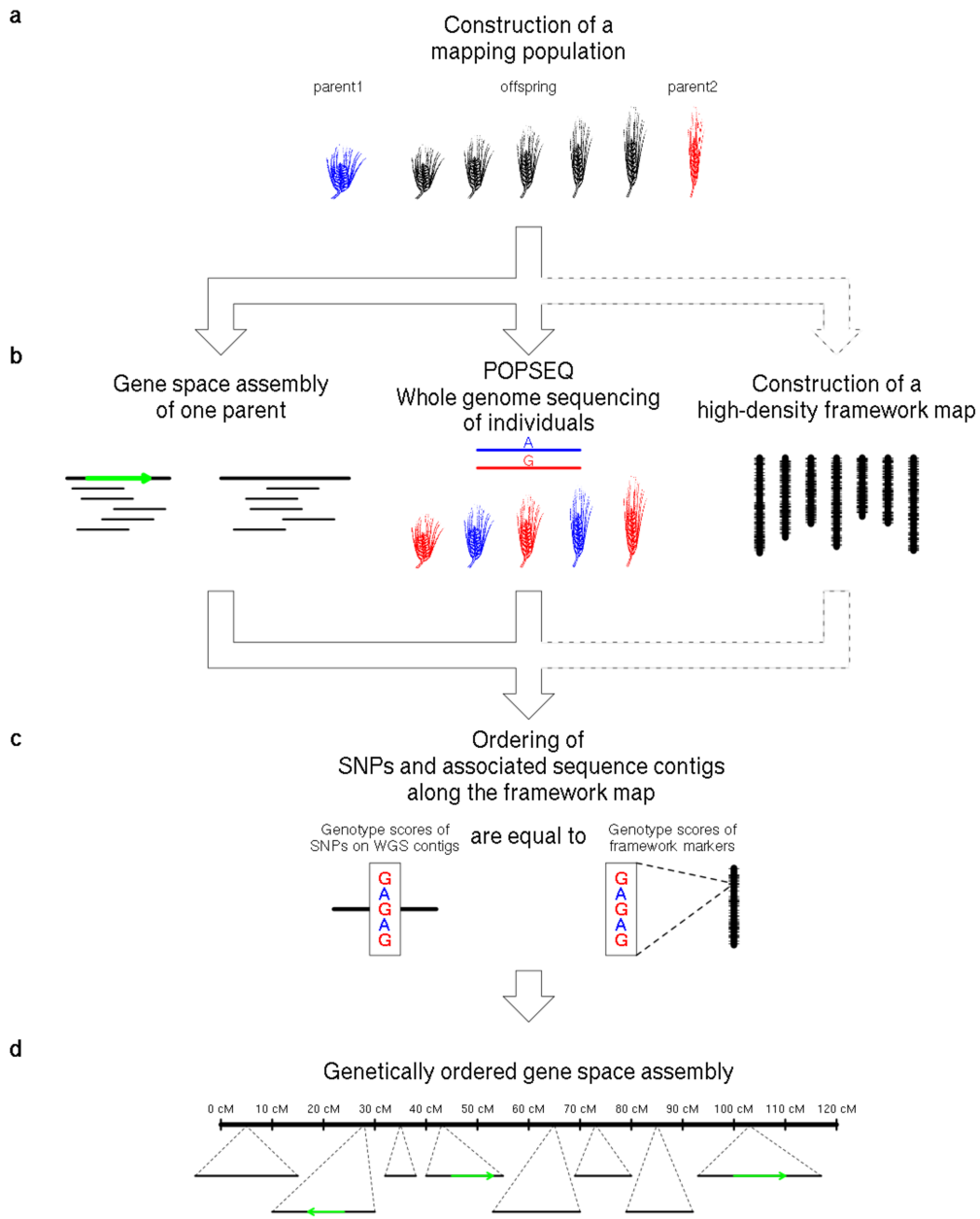
Can we do better? The idea of POPSEQ

- Only 410 Mb could be positioned in the physical framework.
- The number of genetic markers limits anchoring efficiency.
- Next-generation sequencing has been used in rice and fruit fly for genotyping. Marker order was derived from a high quality reference genome.

no. of contigs	2.7 million
cumulative length	1.6 Gb
mean contig length	700 bp
no. contigs $>$ 1kb	376,261
lengh of contig $>$ 1kb	1.1 Gb
N50	1,425 bp

Can we do better? The idea of POPSEQ

- Only 410 Mb could be positioned in the physical framework.
- The number of genetic markers limits anchoring efficiency.
- Next-generation sequencing has been used in rice and fruit fly for genotyping. Marker order was derived from a high quality reference genome.
- Idea: use whole genome sequencing for genotyping to establish marker order from sequencing data



Barley POPSEQ populations

- 90 Morex $\times$ Barke (MxB) RILs
  - Parents of the current barley reference population

Barley POPSEQ populations

- 90 Morex $\times$ Barke (MxB) RILs
  - Parents of the current barley reference population
- 82 Oregon Wolfe Barley (OWB) DH lines
  - progeny from a cross between dominant and recessive marker stocks

from Oregon State University

Sequencing the populations

- WGS resequencing of Morex × Barke and OWB
- Read mapping and SNP calling with BWA/samtools pipeline

	MxB WGS	OWB WGS
Population	Morex × Barke	Oregon Wolfe Barleys
	RIL F8	DH
Seq. technology	WGS; Hiseq 2000	WGS; Hiseq 2000
No. of lanes	12	12
No. of individuals	90 (+parents)	82 (+parents)
Coverage per sample	∼1x	∼1x
No. of SNPs	5.1 M	6.5 M

Putting together the pieces

- Annotated WGS contigs of cultivar 'Morex' are available (IBSC, Nature, 2012)

Putting together the pieces

- Annotated WGS contigs of cultivar 'Morex' are available (IBSC, Nature, 2012)
- A high-density genetic map ("iSelect map") had been constructed through array-based genotyping of 360 MxB RILs (Comadran et al., Nat. Genet., 2012)

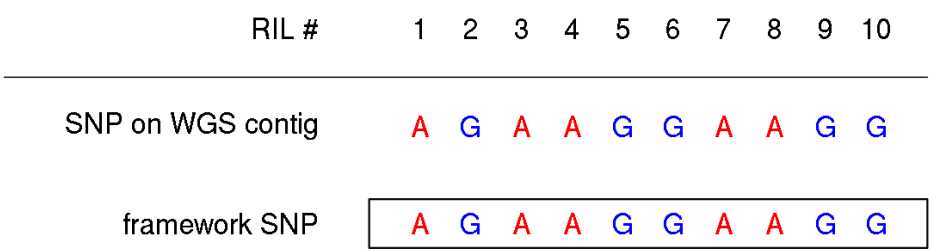
Putting together the pieces

- Annotated WGS contigs of cultivar 'Morex' are available (IBSC, Nature, 2012)
- A high-density genetic map ("iSelect map") had been constructed through array-based genotyping of 360 MxB RILs (Comadran et al., Nat. Genet., 2012)
- SNP and associated WGS contigs were placed into this framework though nearest neighbor search

RIL #	1	2	3	4	5	6	7	8	9	10
SNP on WGS contig	A	G	A	A	G	G	A	A	G	G

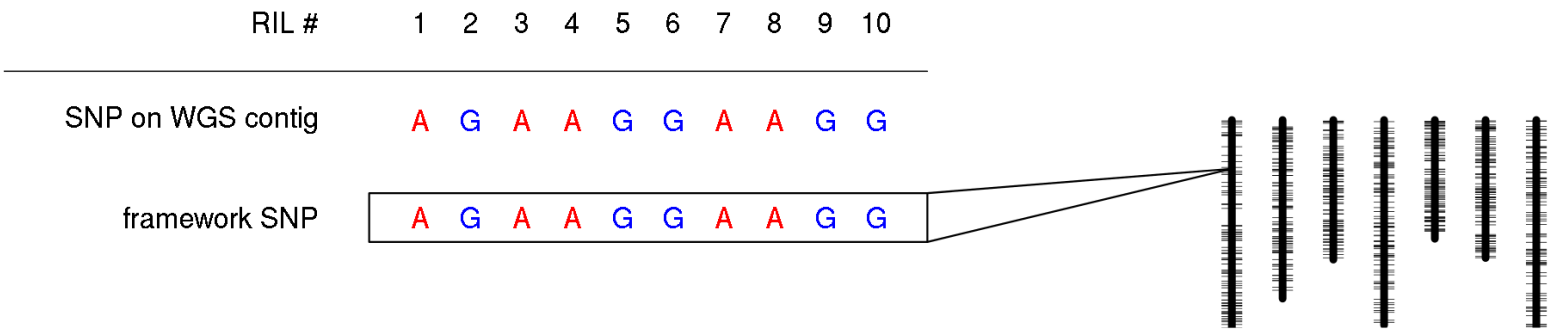
Putting together the pieces

- Annotated WGS contigs of cultivar 'Morex' are available (IBSC, Nature, 2012)
- A high-density genetic map ("iSelect map") had been constructed through array-based genotyping of 360 MxB RILs (Comadran et al., Nat. Genet., 2012)
- SNP and associated WGS contigs were placed into this framework though nearest neighbor search



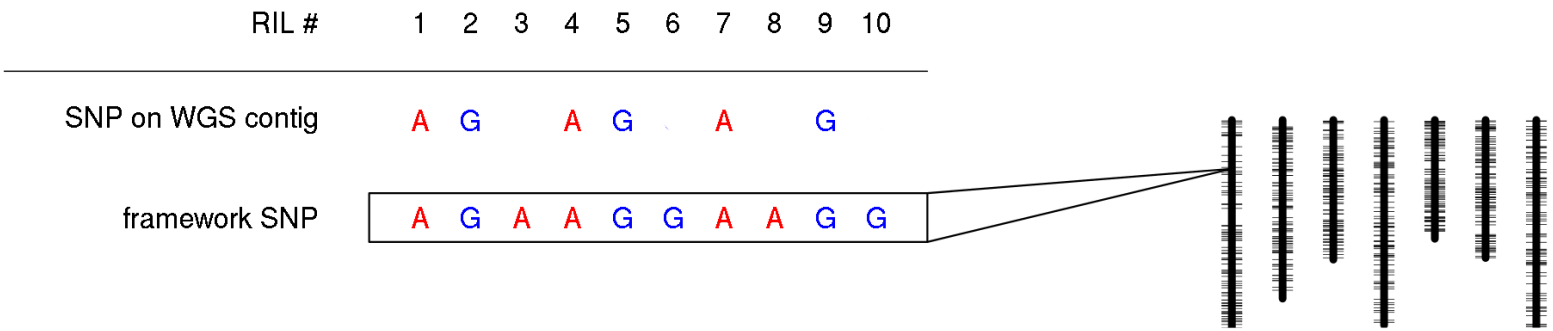
Putting together the pieces

- Annotated WGS contigs of cultivar 'Morex' are available (IBSC, Nature, 2012)
- A high-density genetic map ("iSelect map") had been constructed through array-based genotyping of 360 MxB RILs (Comadran et al., Nat. Genet., 2012)
- SNP and associated WGS contigs were placed into this framework though nearest neighbor search



Putting together the pieces

- Annotated WGS contigs of cultivar 'Morex' are available (IBSC, Nature, 2012)
- A high-density genetic map ("iSelect map") had been constructed through array-based genotyping of 360 MxB RILs (Comadran et al., Nat. Genet., 2012)
- SNP and associated WGS contigs were placed into this framework though nearest neighbor search



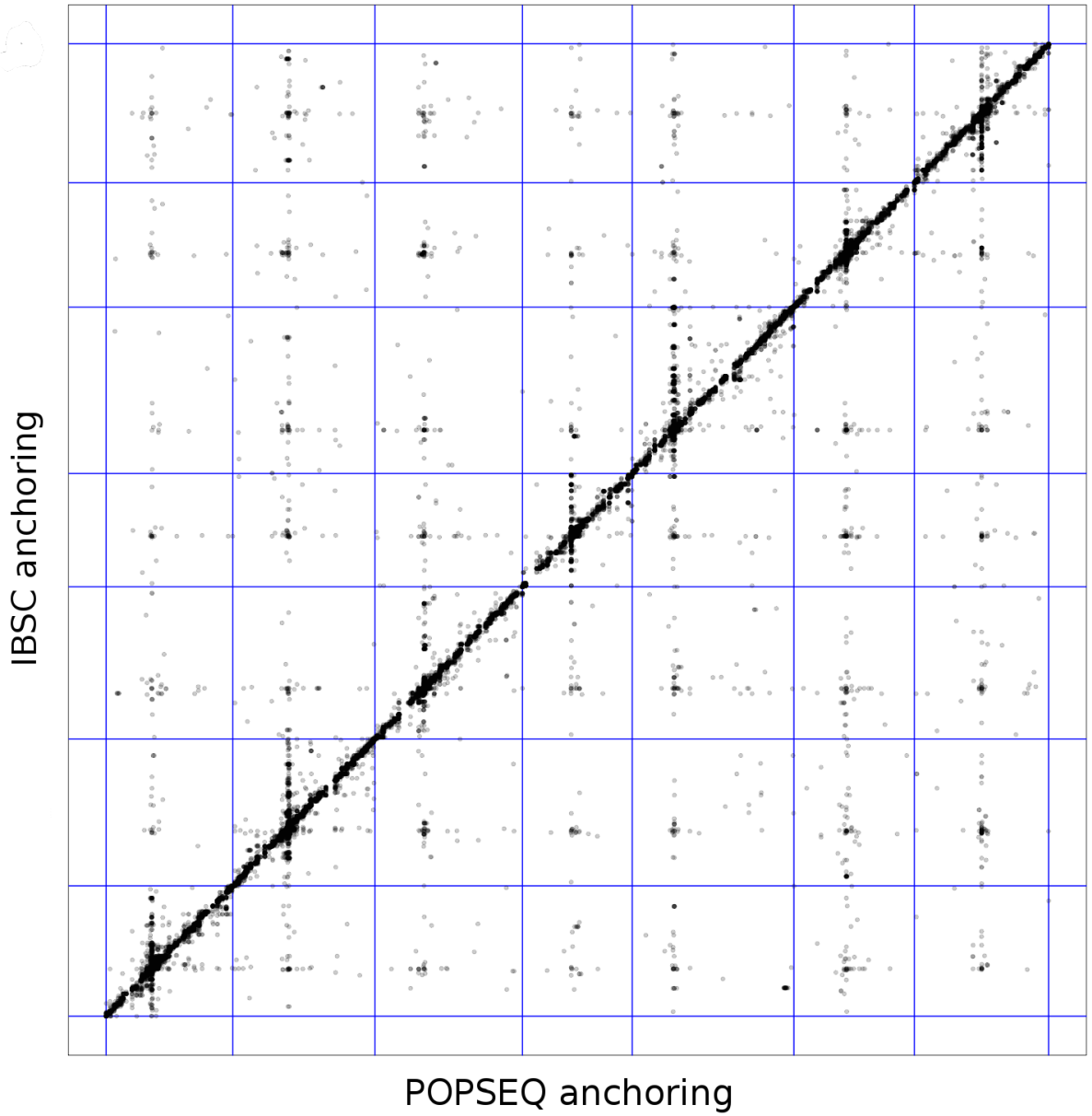
- Consistency criteria when there are multiple nearest framework markers

Anchoring to the Morex × Barke iSelect framework

	MxB WGS	IBSC
Framework map	iSelect	iSelect
No. of SNPs used for anchoring	4,381,020	498,165
No. of anchored contigs	498,856	138,443
Size of anchored contigs	927 Mb	410 Mb
	(50%)	(21%)
No. of anchored HC genes	16,682	14,923
	(64%)	(57%)

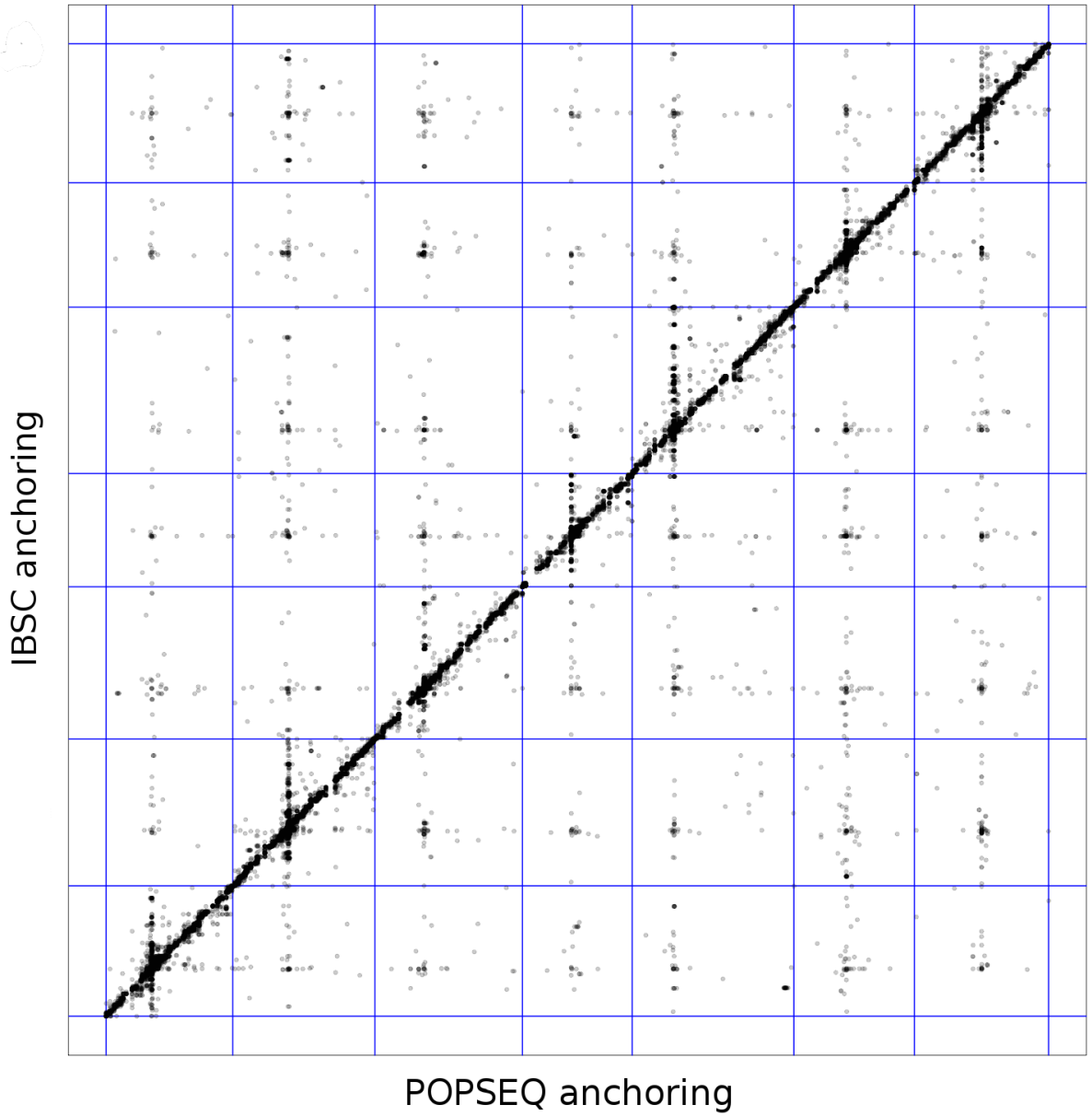
Collinearity between POPSEQ and IBSC anchoring

- 91 % agreement



Collinearity between POPSEQ and IBSC anchoring

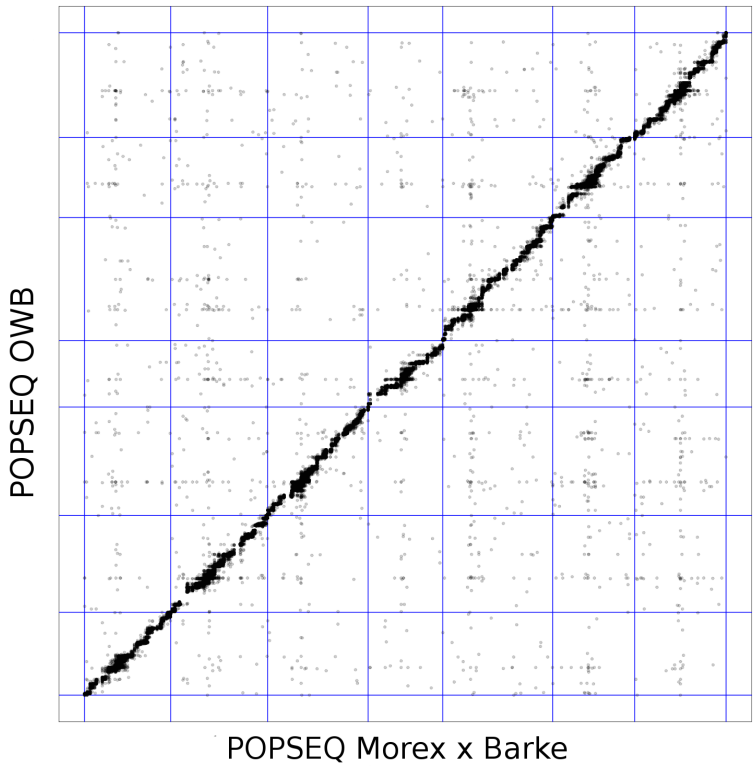
- 91 % agreement
- 95 % of contig pairs on the same BAC are anchored within 3 cM



POPSEQ anchoring to OWB GBS map

	MxB WGS	OWB WGS
No. of SNPs used for anchoring	4,381,020	6,072,994
Framework map	iSelect	OWB GBS
No. of anchored contigs	498,856	584,806
Size of anchored contigs	927 Mb	978 Mb
	(50%)	(52%)
No. of anchored HC genes	16,682	15,171
	(64%)	(58%)

- framework: OWB GBS map
- 93.2 % agreement between maps



Summary

- Combination of Morex × Barke and OWB results to compensate for regions that are non-polymorphic in one population

	MxB + OWB WGS	IBSC
No. of SNPs used for anchoring	11,229,709	498,165
Framework map	iSelect/OWB GBS	iSelect
No. of anchored contigs	747,077	138,443
Size of anchored contigs	1,222 Mb (65%)	410 Mb (21%)
No. of anchored HC genes	20,932 (80%)	14,923 (57%)

- Three times more anchored WGS contigs compared to the physical and genetic framework

What can POPSEQ do for you?

- Genetically anchored gene-space assembly from cheap NGS reads
- No need for physical mapping and long sequence contigs
- Independent of prior genomic resources

What can POPSEQ do for you?

- Genetically anchored gene-space assembly from cheap NGS reads
- No need for physical mapping and long sequence contigs
- Independent of prior genomic resources
- Can be applied to other (crop) species (wheat, rye, orphan crops, wild relatives)




Images from Wikipedia

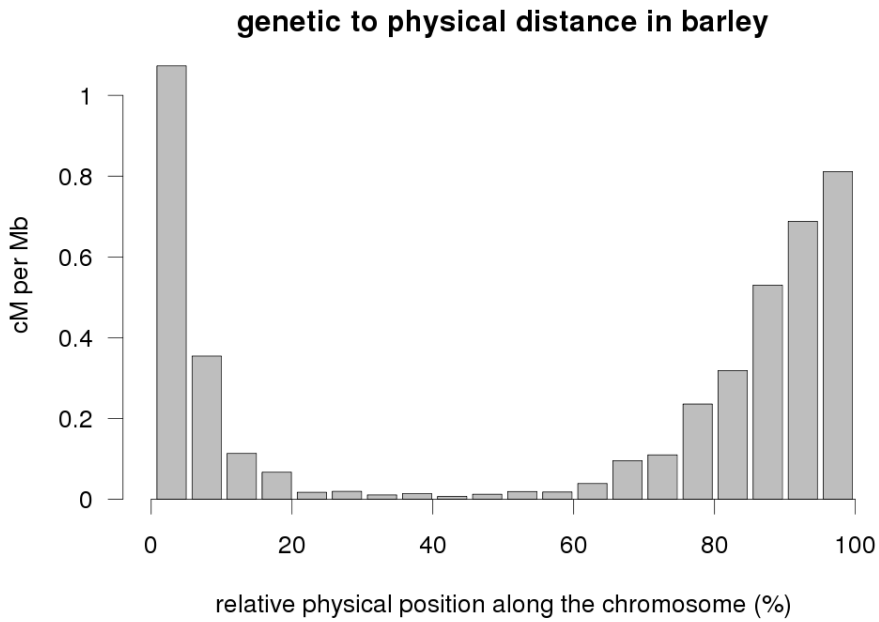
POPSEQ anchoring of the physical map

	BAC contigs	sequenced clones
POPseq data	MxB + OWB	MxB + OWB
# all contigs	9,265	6,278
# with WGS contigs	5,872	6,243
# with anc. WGS contigs	5,720	6,189
# anchored	5,193	5,591
length	3.95 Gb	703 Mb

- POPSEQ can assign additional physical contigs to chromosomes to assist MTP sequencing

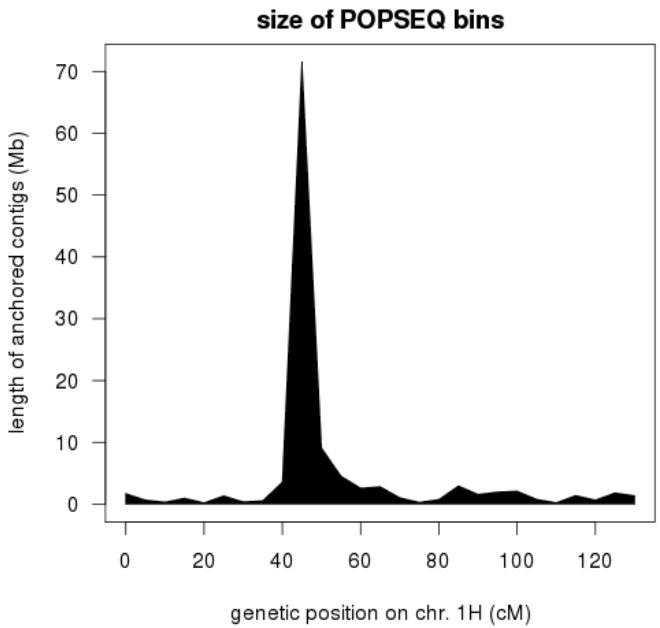
Challenges and limitations

- POPSEQ relies on recombination



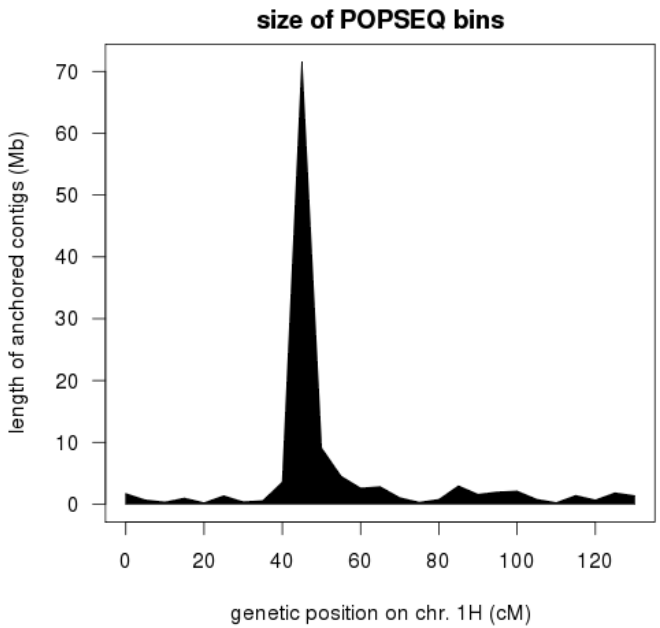
Challenges and limitations

- POPSEQ relies on recombination



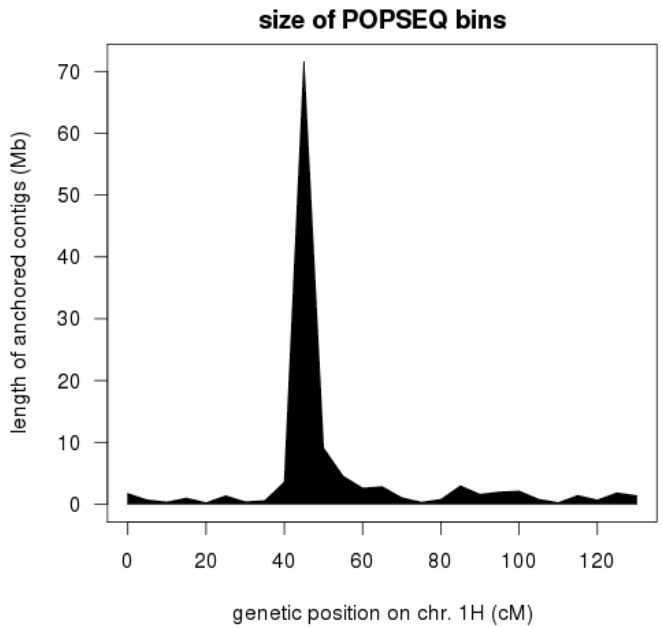
Challenges and limitations

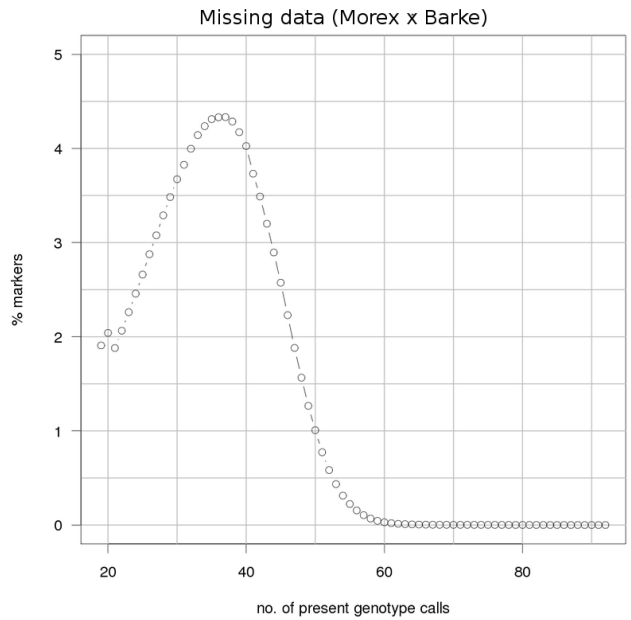
- POPSEQ relies on recombination
- Assembly quality (contig size and number)



Challenges and limitations

- POPSEQ relies on recombination
- Assembly quality (contig size and number)
- Current sequencing costs limit sequencing depth, population size and mapping resolution





Acknowledgements

- Nils Stein
- Uwe Scholz
- Axel Himmelbach
- Ruvini Ariyadasa
- Robbie Waugh
- Gary Muehlbauer
- Jesse Poland
- Dan Rokhsar
- Jarrod Chapman
- Jeremy Schmutz
- Kerrie Barry
- María Muñoz-Amatríain
- Klaus Mayer
- Alan Schulman
- Tim Close
- Roger Wise