Genetic anchoring of the chromosome shotgun assembly of bread wheat by population sequencing

Martin Mascher

IPK Gatersleben

EUCARPIA ITMI meeting

June 29, 2014

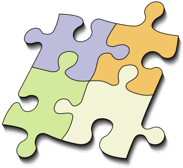
Member of the Leibniz Leibniz Association

IPK GATERSLEBEN

Shotgun assemblies and why we need to anchor them

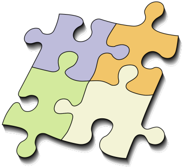
- Whole genome or chromosome shotgun assemblies are a fast and easy way to generate genomic resources.

Shotgun assemblies and why we need to anchor them



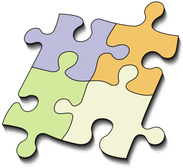
- Whole genome or chromosome shotgun assemblies are a fast and easy way to generate genomic resources.
- Shotgun assemblies are half-solved jigsaw puzzles.

Shotgun assemblies and why we need to anchor them



- Whole genome or chromosome shotgun assemblies are a fast and easy way to generate genomic resources.
- Shotgun assemblies are half-solved jigsaw puzzles.
- Genetic mapping to assign assembly contigs to chromosomal locations

Shotgun assemblies and why we need to anchor them



- Whole genome or chromosome shotgun assemblies are a fast and easy way to generate genomic resources.
- Shotgun assemblies are half-solved jigsaw puzzles.
- Genetic mapping to assign assembly contigs to chromosomal locations
- POPSEQ in barley and wheat

The idea of POPSEQ

- Only 25 % of the barley WGS assembly could be positioned in the physical framework.

no. of contigs	2.7 million
cumulative length	1.8 Gb
mean contig length	700 bp
no. contigs > 1kb	376,261
length of contigs > 1kb	1.1 Gb
N50	1,425 bp

The idea of POPSEQ

- Only 25 % of the barley WGS assembly could be positioned in the physical framework.
- The number of genetic markers limits anchoring efficiency.

no. of contigs	2.7 million
cumulative length	1.8 Gb
mean contig length	700 bp
no. contigs $>$ 1kb	376,261
length of contigs $>$ 1kb	1.1 Gb
N50	1,425 bp

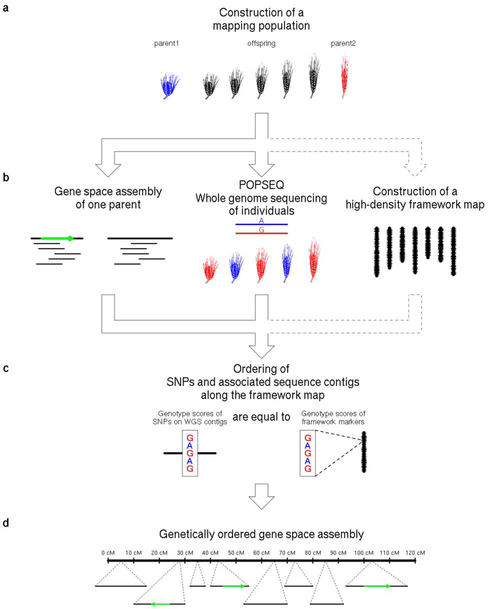
The idea of POPSEQ

- Only 25 % of the barley WGS assembly 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	1.8 Gb
mean contig length	700 bp
no. contigs > 1kb	376,261
length of contigs > 1kb	1.1 Gb
N50	1,425 bp

The idea of POPSEQ

- Only 25 % of the barley WGS assembly 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


a
Construction of a mapping population
parent1
offspring
parent2
b
Gene space assembly of one parent
POPSEQ
Whole genome sequencing of individuals
A
G
Construction of a high-density framework map
c
Ordering of SNPs and associated sequence contigs along the framework map
Genotype scores of SNPs on WGS contigs
are equal to
Genotype scores of framework markers
G A G A G
G A G A G
d
Genetically ordered gene space assembly
0 cM
10 cM
20 cM
30 cM
40 cM
50 cM
60 cM
70 cM
80 cM
90 cM
100 cM
110 cM
120 cM

POPSEQ results in barley

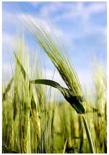
- POPSEQ was done with one RIL (Morex × Barke) and one DH population (OWB).

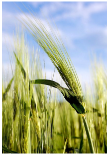
	MxB + OWB WGS	IBSC
No. of SNPs used for anchoring	11,229,709	498,165
Framework map	iSelect/OWB GBS	iSelect
No. of anchored contigs	747,077	138,443
Size of anchored contigs	1,222 Mb (65%)	410 Mb (21%)
Median length of anchored contigs	891 bp	1,775 bp
No. of anchored HC genes	20,932 (80%)	14,923 (57%)

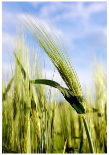
- Three times more anchored sequence compared to the physical and genetic framework

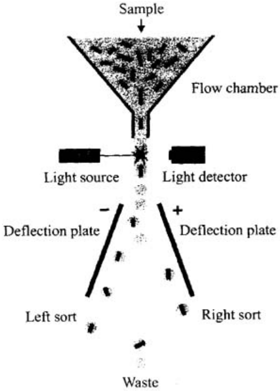
Chromosome shotgun sequencing (CSS) in wheat

=



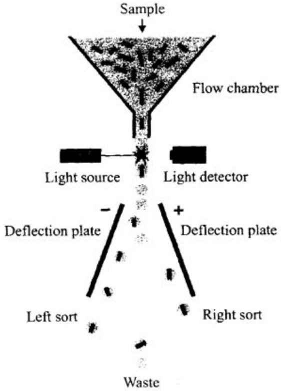






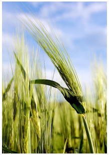
Chromosome shotgun sequencing (CSS) in wheat

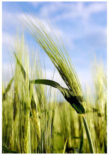
- IWGSC has created shotgun sequence assemblies of all 40 wheat chromosome arms + 3B
- Single chromosome arms were isolated from cytogenetic stocks using flow cytometry

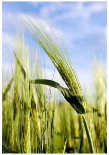


Chromosome shotgun sequencing (CSS) in wheat

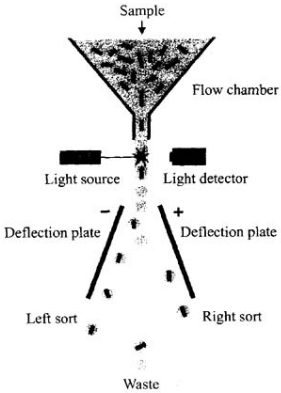
=







- IWGSC has created shotgun sequence assemblies of all 40 wheat chromosome arms + 3B
- Single chromosome arms were isolated from cytogenetic stocks using flow cytometry
- DNA libraries of sorted chromosomes were sequenced to high coverage on the HiSeq2000 and assembled by TGAC
- Total assembly size: 10.1 Gb



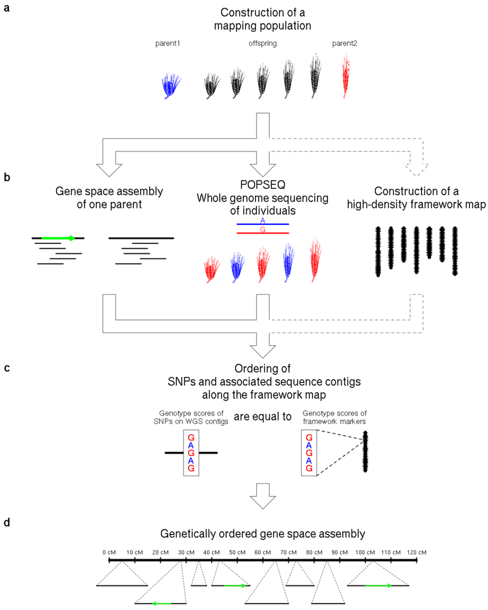
Sequencing the Synthetic W7984 × Opata M85 population

- POPSEQ anchoring of the CSS assembly by sequencing the SynOp doubled haploid population
- Synthetic wheat: artificial hybridization of a tetraploid durum wheat with *Ae. tauschii*.

Sequencing the Synthetic W7984 × Opata M85 population

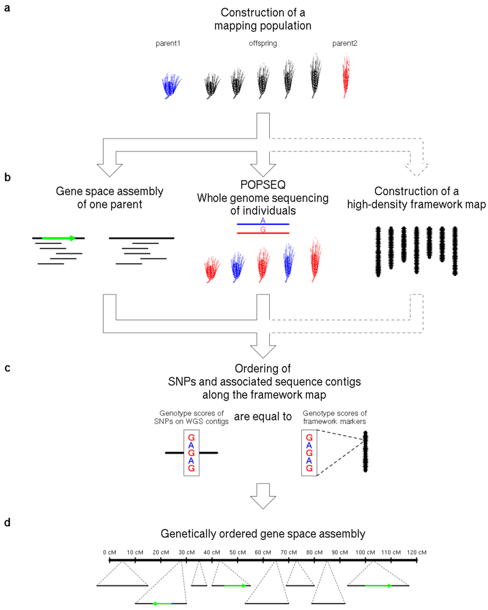
- POPSEQ anchoring of the CSS assembly by sequencing the SynOp doubled haploid population
- Synthetic wheat: artificial hybridization of a tetraploid durum wheat with *Ae. tauschii*.
- JGI sequenced 90 doubled haploid lines to 1x coverage.
- Read mapping and SNP calling were done with BWA and SAMTools.

POPSEQ: putting together the pieces


a
Construction of a mapping population
parent1
offspring
parent2
b
Gene space assembly of one parent
POPSEQ Whole genome sequencing of individuals
A
G
Construction of a high-density framework map
c
Ordering of SNPs and associated sequence contigs along the framework map
Genotype scores of SNPs on WGS contigs
are equal to
Genotype scores of framework markers
G A G A G
G A G A G
d
Genetically ordered gene space assembly
0 cM
10 cM
20 cM
30 cM
40 cM
50 cM
60 cM
70 cM
80 cM
90 cM
100 cM
110 cM
120 cM

- Annotated sequence contigs of the wheat CSS assembly (IWGSC)

POPSEQ: putting together the pieces



a
Construction of a mapping population
parent1
offspring
parent2
b
Gene space assembly of one parent
POPSEQ Whole genome sequencing of individuals
A
G
Construction of a high-density framework map
c
Ordering of SNPs and associated sequence contigs along the framework map
Genotype scores of SNPs on WGS contigs
are equal to
Genotype scores of framework markers
G A G A G
G A G A G
d
Genetically ordered gene space assembly
0 cM
10 cM
20 cM
30 cM
40 cM
50 cM
60 cM
70 cM
80 cM
90 cM
100 cM
110 cM
120 cM

- Annotated sequence contigs of the wheat CSS assembly (IWGSC)
- A high-density genetic map was constructed through GBS of the Synthetic x Opata population (Poland, 2012).

Placing the SNPs into a framework map

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

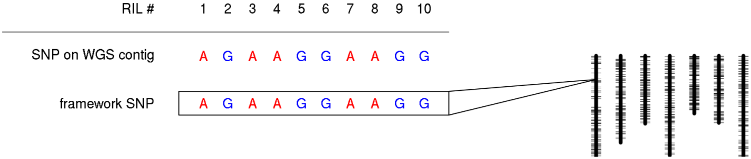
- WGS SNPs and framework markers are represented as binary genotype vectors.

Placing the SNPs into a framework map

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
framework SNP	A	G	A	A	G	G	A	A	G	G

- WGS SNPs and framework markers are represented as binary genotype vectors.
- The nearest neighbor(s) (Hamming distance) are searched for in the set of framework markers whose genetic positions are known.

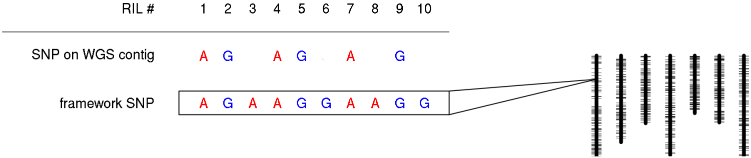
Placing the SNPs into a framework map



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
framework SNP	A	G	A	A	G	G	A	A	G	G

- WGS SNPs and framework markers are represented as binary genotype vectors.
- The nearest neighbor(s) (Hamming distance) are searched for in the set of framework markers whose genetic positions are known.

Placing the SNPs into a framework map



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

- WGS SNPs and framework markers are represented as binary genotype vectors.
- The nearest neighbor(s) (Hamming distance) are searched for in the set of framework markers whose genetic positions are known.
- Consistency criteria for multiple nearest neighbors: framework are required to be within 5 cM.

Wheat POPSEQ results

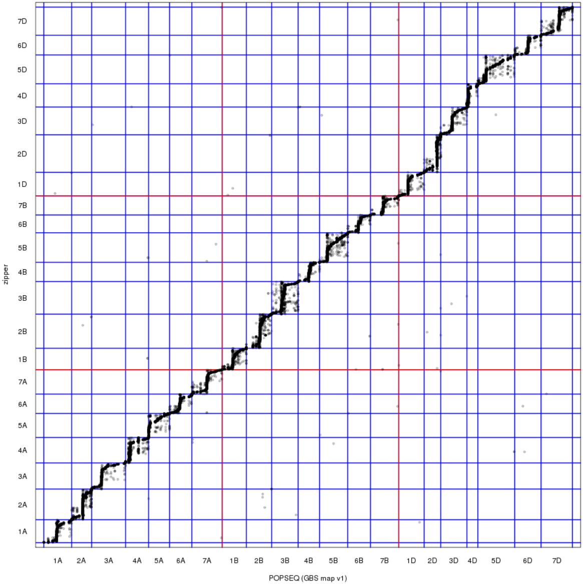
	wheat	barley
population	SynOp DH	OWB DH
assembly size	10.1 Gb (63 %)	1.8 Gb (38 %)
N50	2,308 bp	1,425 bp
size in contigs $\geq$ 1 kb	7.0 Gb	1.1Gb
size in contigs $\geq$ 5 kb	3.1 Gb	382 Mb
anchored length	4.4 Gb	1.0 Gb
anchored length $\geq$ 1 kb	4.2 Gb	811 Mb
anchored length $\geq$ 5 kb	2.3 Gb	279 Mb

Wheat POPSEQ results

	wheat	barley
population	SynOp DH	OWB DH
assembly size	10.1 Gb (63 %)	1.8 Gb (38 %)
N50	2,308 bp	1,425 bp
size in contigs $\geq$ 1 kb	7.0 Gb	1.1Gb
size in contigs $\geq$ 5 kb	3.1 Gb	382 Mb
anchored length	4.4 Gb	1.0 Gb
anchored length $\geq$ 1 kb	4.2 Gb	811 Mb
anchored length $\geq$ 5 kb	2.3 Gb	279 Mb

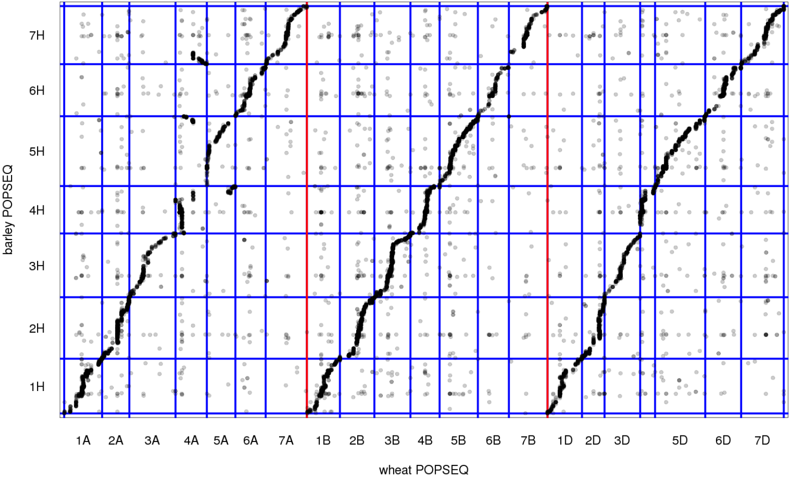
- 99.4 % agreement between POPSEQ and flow sorting

Collinearity with the GenomeZipper



- 99.8 % agreement of chromsome assignments
- 85 % correlation within linkage groups
- 75,183 genes anchored by POPSEQ and/or GenomeZipper

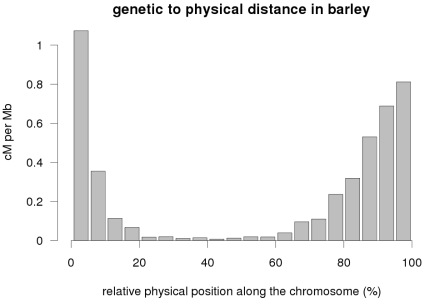
Collinearity with barley



- 93 % agreement of group assignments; 91 % collinearity within groups

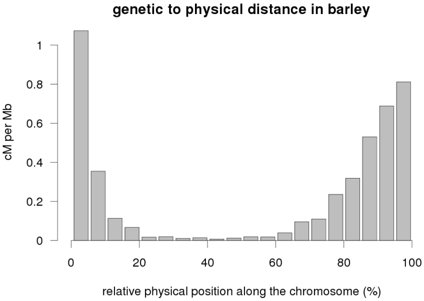
Challenges and limitations of POPSEQ

- Biology: POPSEQ relies on recombination.



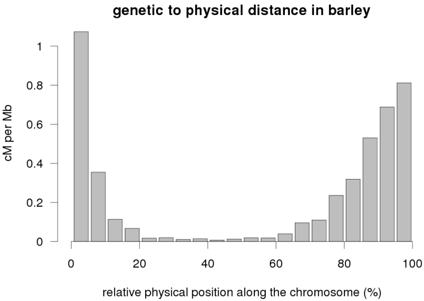
Challenges and limitations of POPSEQ

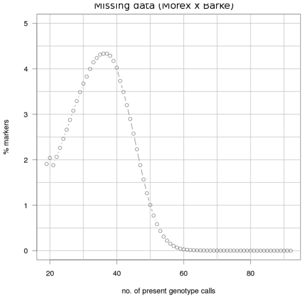
- Biology: POPSEQ relies on recombination.
- Algorithms: assembly quality (contig size and number)



Challenges and limitations of POPSEQ

- Biology: POPSEQ relies on recombination.
- Algorithms: assembly quality (contig size and number)
- Technology/money: sequencing costs limit sequencing depth, population size and mapping resolution.





Acknowledgements

- Nils Stein
- Uwe Scholz
- IWGSC
- Dan Rokhsar
- Jarrod Chapman
- Kerrie Barry
- Robbie Waugh
- Jesse Poland
- Gary Muehlbauer

IPK GATERSLEBEN

PLANT 2030

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

International
Wheat Genome
Sequencing
Consortium

JGI

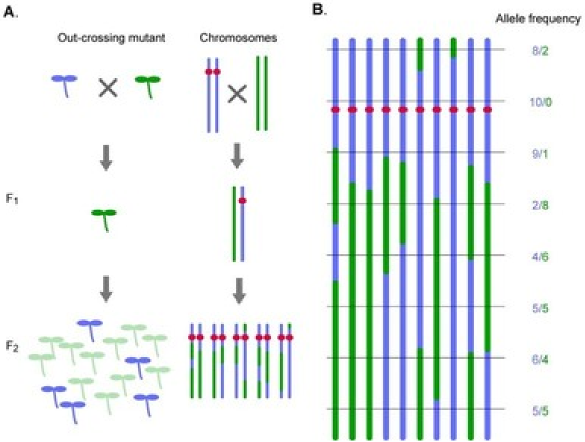
KANSAS STATE
UNIVERSITY

The James
Hutton
Institute

UNIVERSITY
OF MINNESOTA
Driven to Discover℠

Mapping-by-sequencing

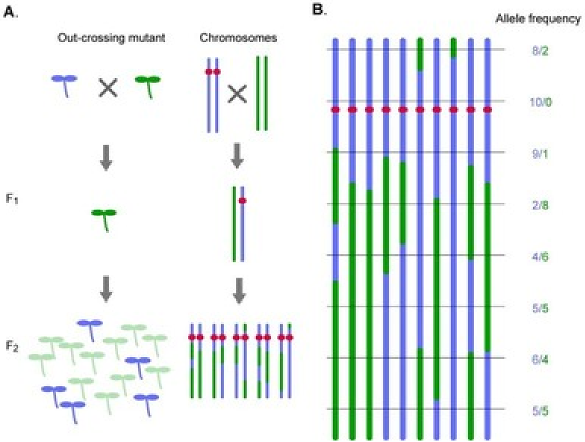
- Identification of causal genes by sequencing phenotypic pools



Schneeberger *et al.*, TIPS 2010

Mapping-by-sequencing

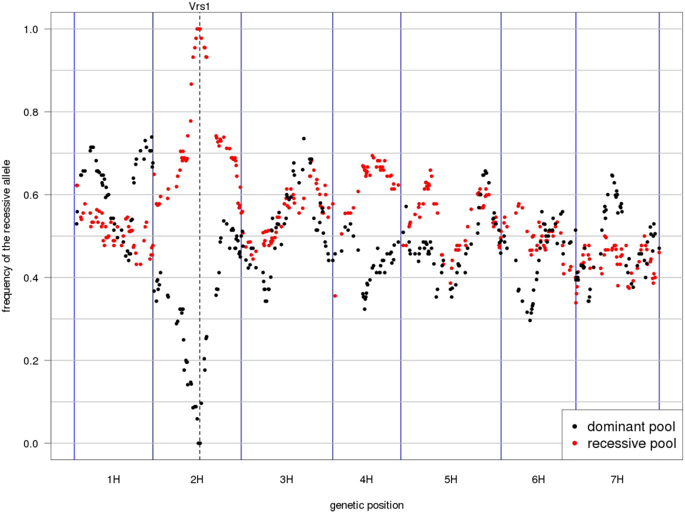
- Identification of causal genes by sequencing phenotypic pools
- Requires an ordered reference sequence



Schneeberger *et al.*, TIPS 2010

Mapping-by-sequencing

- Mapping-by-sequencing of the sixed-row spike gene (vrs1) in OWB



POPSEQ anchoring of the barley 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