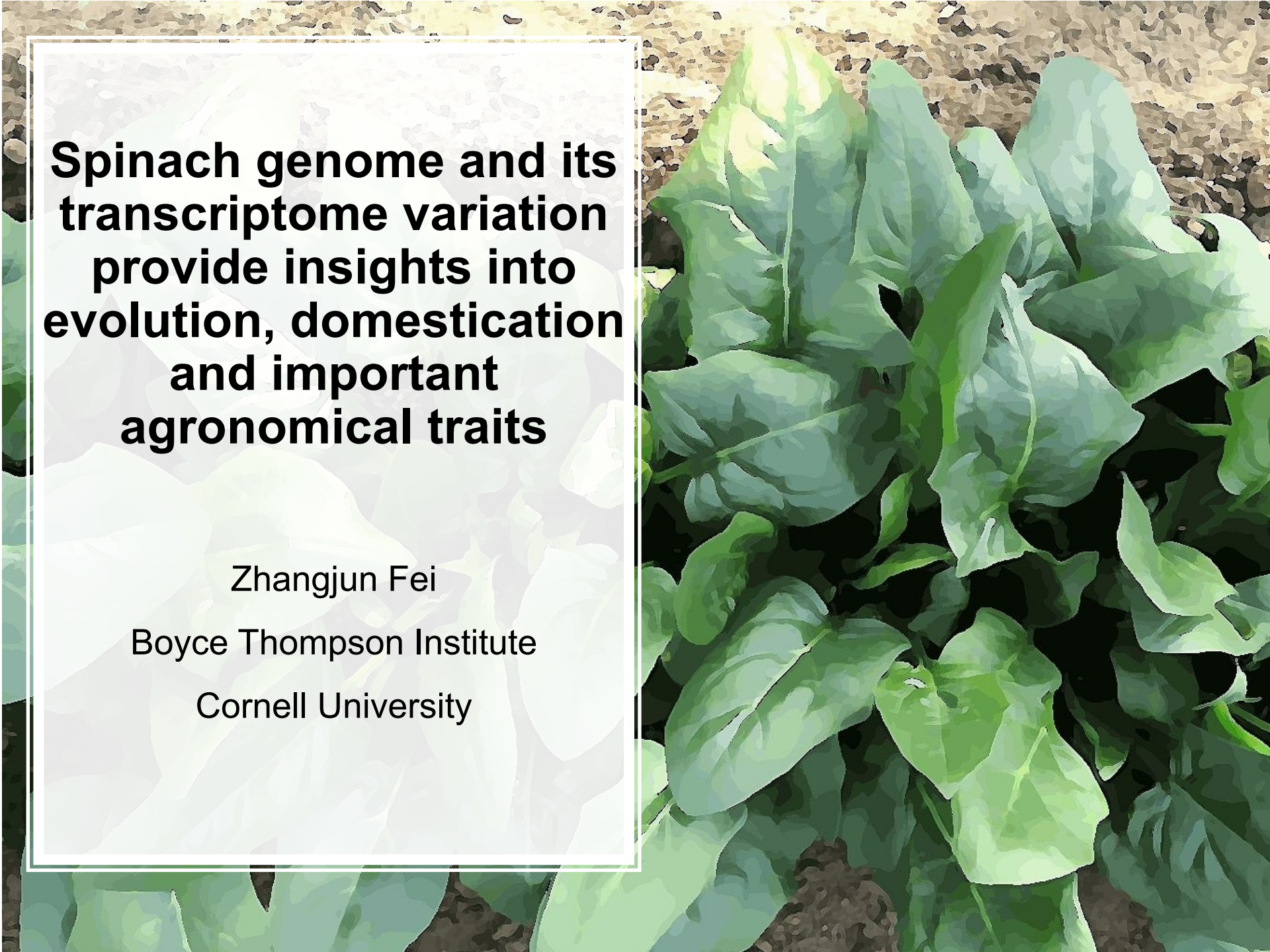
Spinach genome and its transcriptome variation provide insights into evolution, domestication and important agronomical traits

Zhangjun Fei

Boyce Thompson Institute

Cornell University



Importance

One of the most nutritious food

Proximates			Minerals			Vitamins		
	per 100 g	Unit		per 100 g	Unit		per 100 g	Unit
Water	91.4	g	Calcium, Ca	99	mg	Vitamin C	28.1	mg
Energy	23	kcal	Iron, Fe	2.71	mg	Thiamin	0.078	mg
Protein	2.86	g	Magnesium, Mg	79	mg	Riboflavin	0.189	mg
Total lipid	0.39	g	Phosphorus, P	49	mg	Niacin	0.724	mg
Carbohydrate	3.63	g	Potassium, K	558	mg	Vitamin B-6	0.195	mg
Fiber	2.2	g	Sodium, Na	79	mg	Folate, DFE	194	µg
Sugars	0.42	g	Zinc, Zn	0.53	mg	Vitamin A, RAE	469	µg
						Vitamin A, IU	9377	IU
						Vitamin E	2.03	mg
						Vitamin K	482.9	µg

https://ndb.nal.usda.gov/

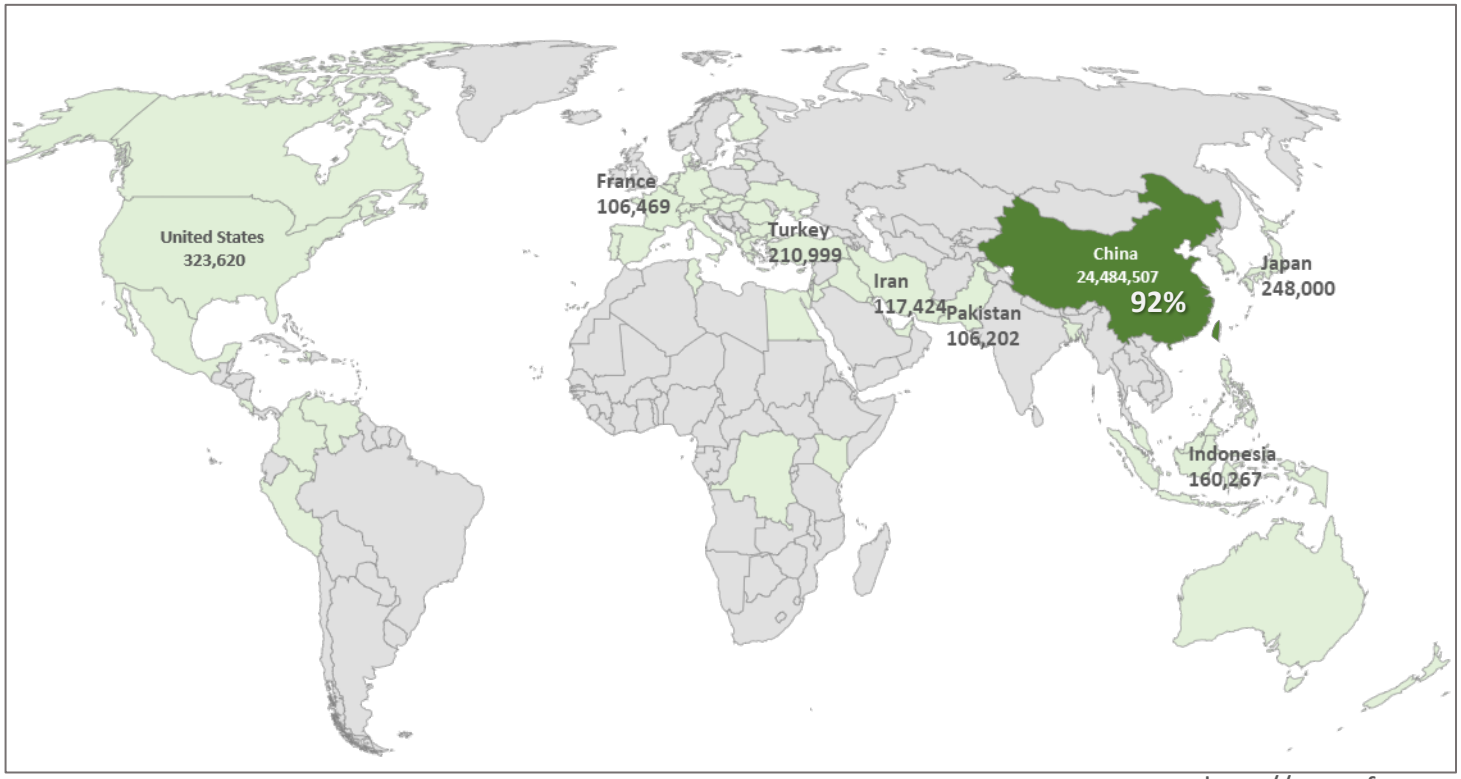
http://www.refoxrelocation.com

http://www.foodnetwork.com/

Production

Planted in ~ 60 countries with an annual production of ~27 Million tones

Spinach production 2016 (tonnes)



http://www.fao.org

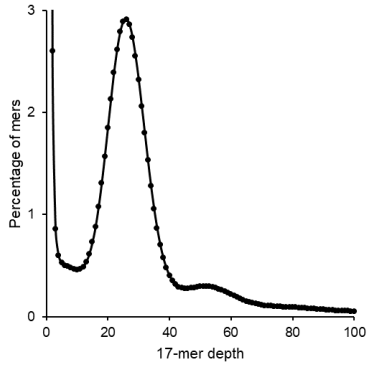
Origin and genetic resource

- Native to central Asia and thought to have originated in Persia (Ryder, 1979).
- Very limited genetic stocks, especially wild accessions.
- Limited genetic and genomic resources

National Plant Germplasm System	Available	Not Available/inactive
Spinacia oleracea	373	244
Spinacia tetrandra	6	9
Spinacia turkestanica	8	4
Spinacia hybr.	1	1

Genome assembly

- A sibling inbreed spinach line (**Sp75**) was used for de novo genome sequencing
- Summary of spinach genome sequencing data
  - Paired-end libraries: 142 Gb ~141x coverage (150 bp, 200 bp, 300 bp, 500 bp and 1 kb)
  - Mate-pair libraries: 27 Gb ~27x coverage (3 kb, 10 kb and 15 kb)
- BioNano genome map
  - 808,135 molecules (>150 kb) with a total length of 214.9 Gb, representing ~213× coverage



Est. genome size: 1,009 Mb

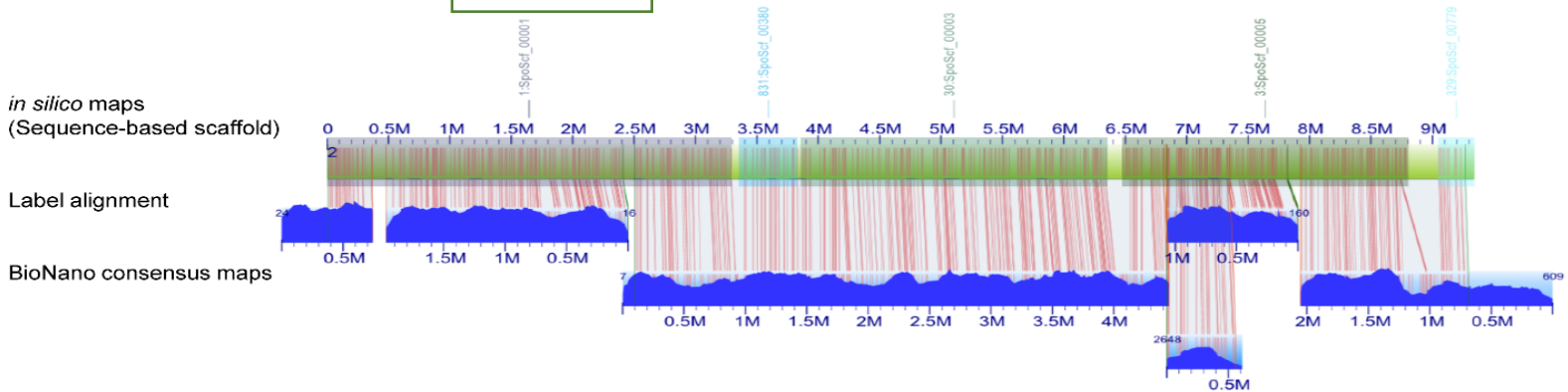
	Contig		Scaffold		Super Scaffold	
	Size (bp)	Number	Size (bp)	Number	Size (bp)	Number
N90	1,554	71,235	5,121	6,093	11,883	3,878
N50	16,570	13,759	319,471	711	919,290	201
N25	31,281	4,483	626,780	218	3,106,702	51
Longest	185,618	1	3,292,865	1	9,343,782	1
Total	830,856,911	215,350	869,796,885	78,264	996,306,834	77,702
			86%			

BUSCO analysis

95.7% of the core eukaryotic genes were completely covered in the genome.

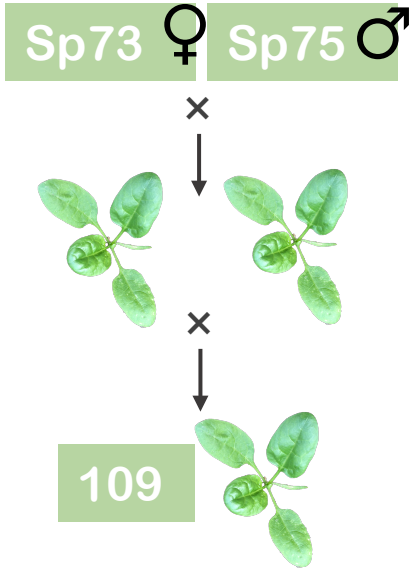
RNA-Seq reads

~95% of the reads could be mapped to the genome.



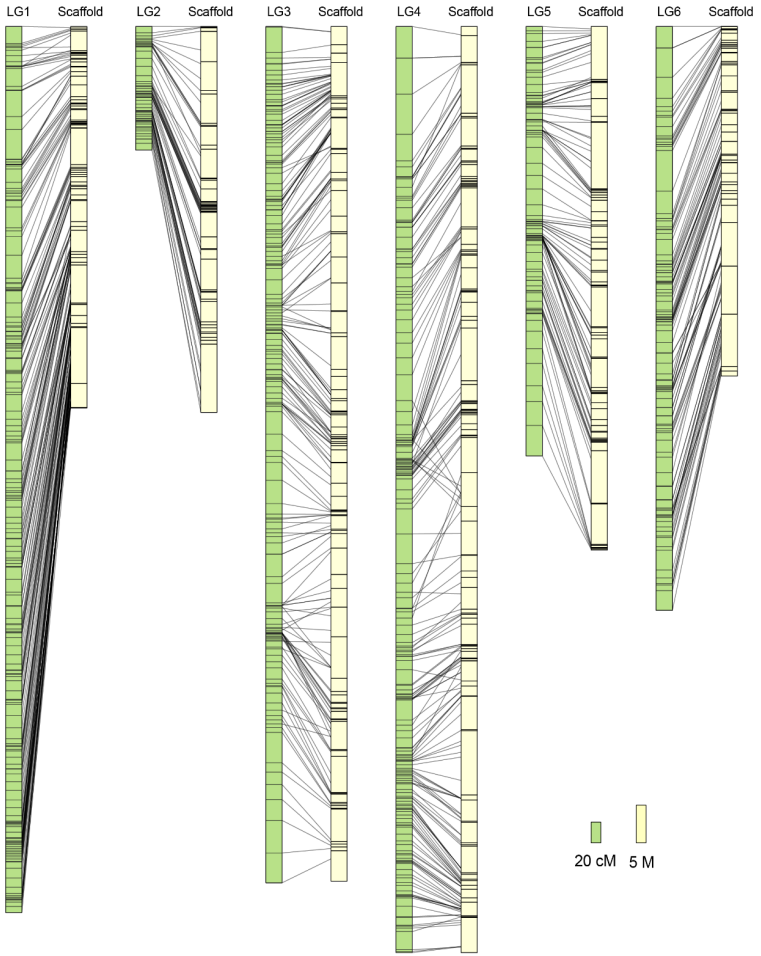
Genetic map

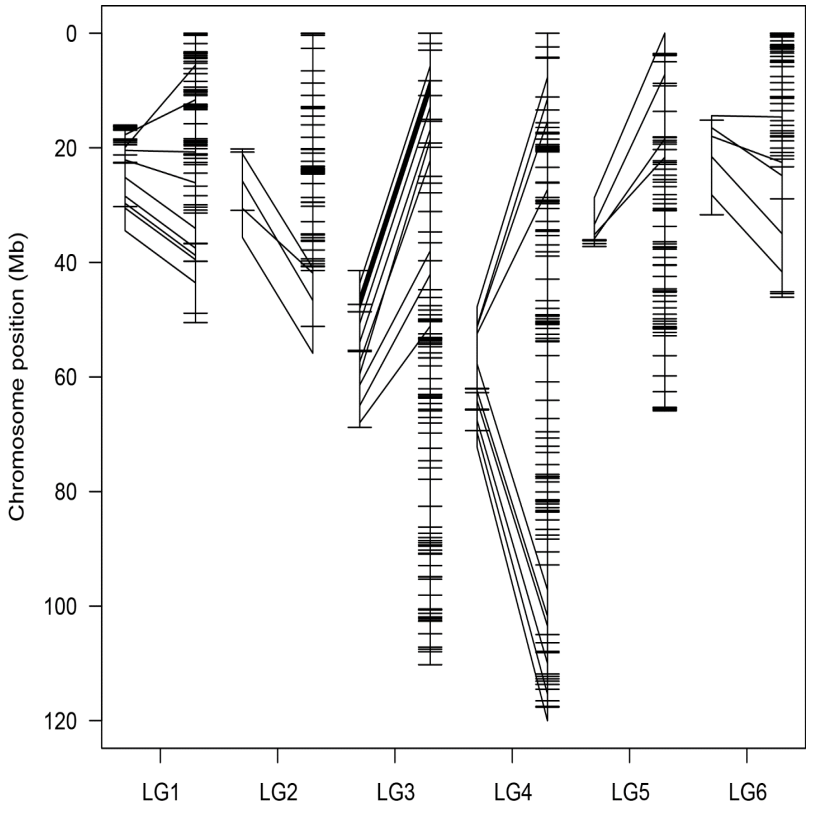
- 109 F2 individuals, were generated by crossing a pair of F1 siblings.
- Genotype by sequencing (GBS)
- 21,792 biallelic SNPs $\xrightarrow{1:2:1}$ 870 SNPs (14%)
- The resulting genetic map consisted of six linkage groups (LGs), corresponding to the six spinach chromosomes. The LGs had an estimated total genetic length of 3679.4 cM and an average of approximately 4.23 cM per marker



Scaffold anchoring

439 scaffolds (47%) were anchored to the six spinach LGs.



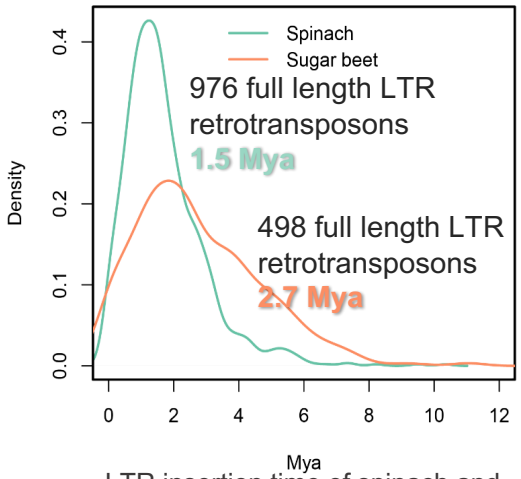


Chan-Navarrete et al., 2016 **(left)** and our map **(right).**

Genome annotation

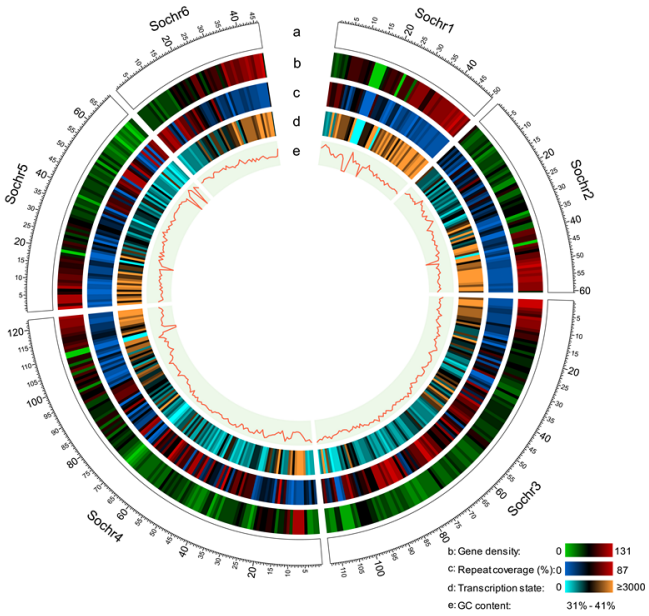
- Repeat contents: 74.4% of the genome (Sugar beet, 42%, 760 Mb)

Class		Family	Count	Masked (bp)	Masked (%)
TEs	Class I	LTR	492,208	470,120,890	**56.58**
		LINE	37,396	23,935,534	3
		SINE	2,181	407,297	0
	Class II	DNA	141,302	50,212,037	6
		MITE	102,617	24,175,497	2.91
		Helitron	1,549	897,226	0.11
	Unknown		144,131	43,656,767	5.25
Satellite			1,143	313,648	0.04
Simple repeat			14,346	4,709,928	0.57
Total			936,873	618,428,824	**74.43**



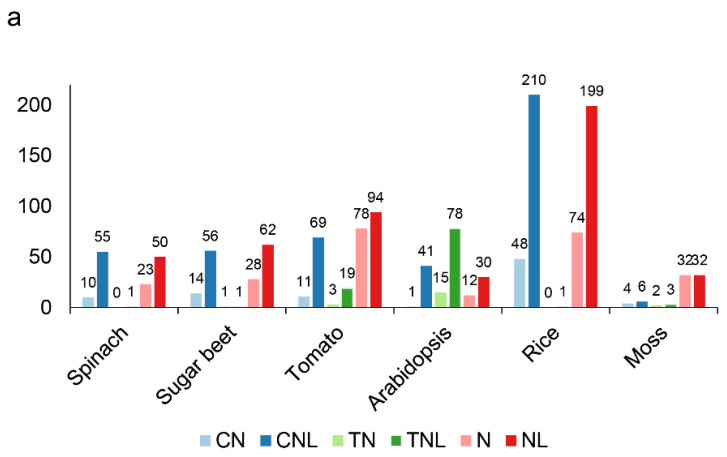
LTR insertion time of spinach and sugar beet.

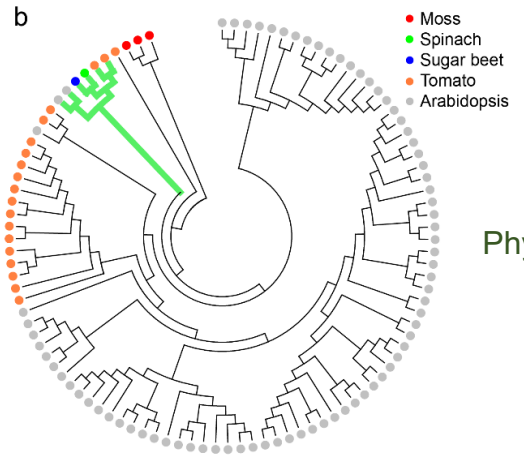
- Protein-coding genes: 25,495
  - 86.15% of the proteins have homologues in nr database.
  - Identified 1202 transcription factors and 892 protein kinases.



R gene and disease resistance

- 139 NBS genes in the spinach genome



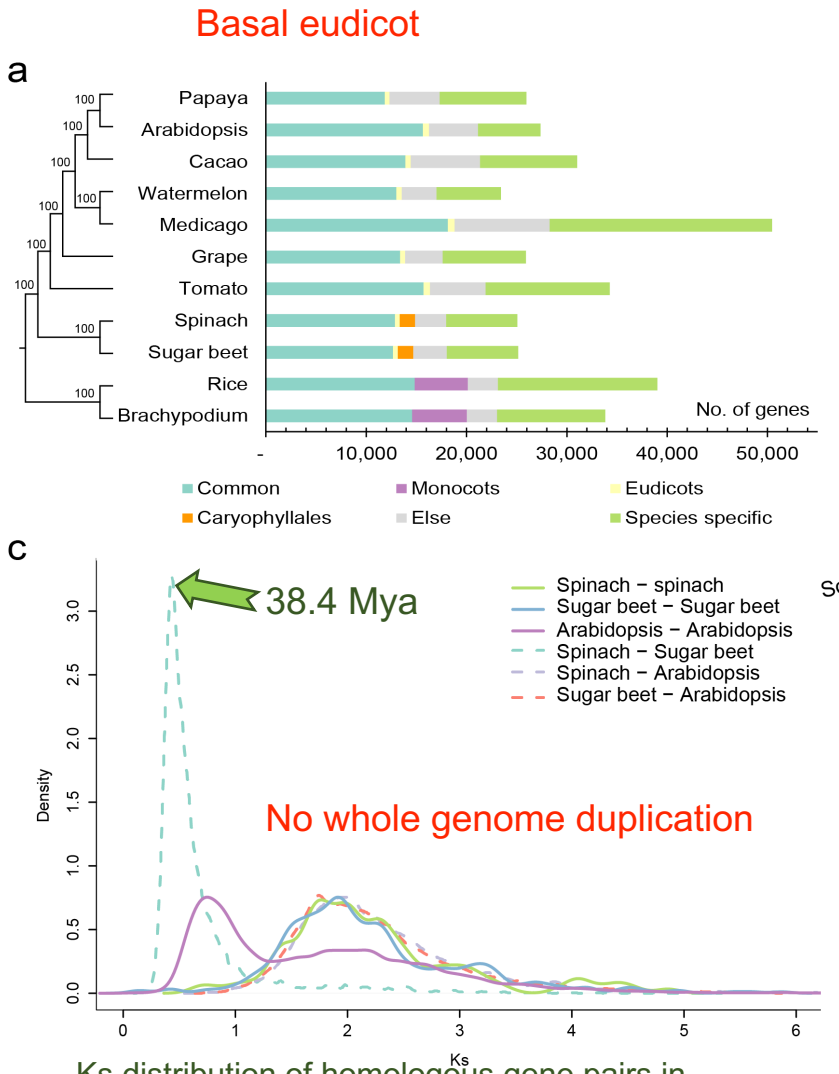


Phylogenetic tree of TNLs

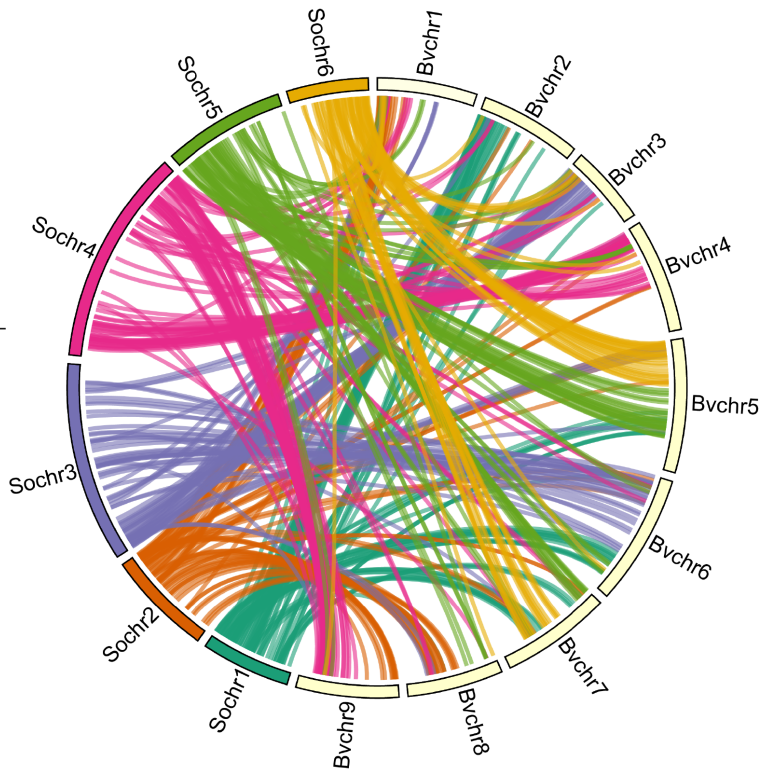
- Resistance against downy mildew pathogen

R1-R5 correspond to R genes Spo12736, Spo12784, Spo12903, Spo12905 and Spo12821, respectively

Comparative genomic analysis



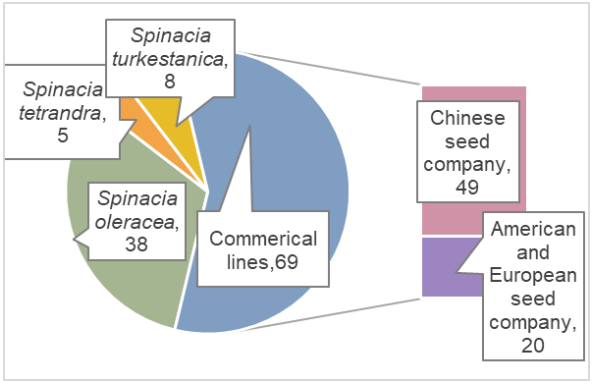
Ks distribution of homologous gene pairs in spinach, sugar beet and Arabidopsis

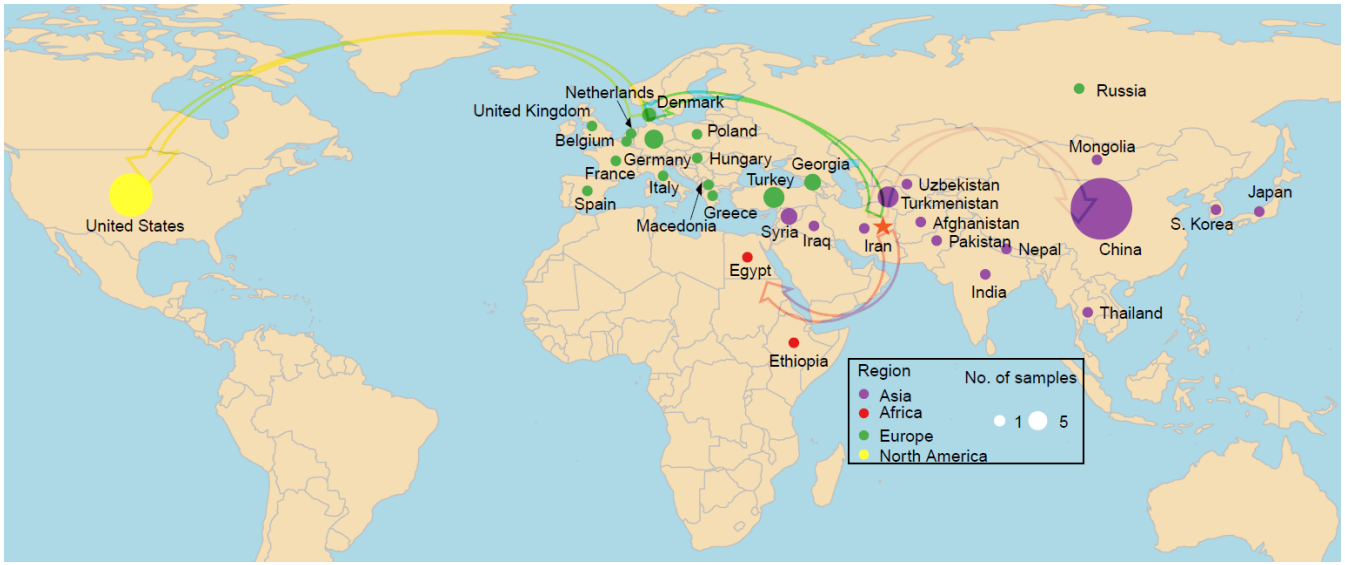


- genomes of spinach and sugar beet shared many syntenic regions
- substantial inter- and intra-chromosome rearrangements during the Caryophyllales genome evolution

Transcriptome diversity

Transcriptome sequencing of 120 cultivated and wild *Spinacia* accessions





Transcriptome diversity

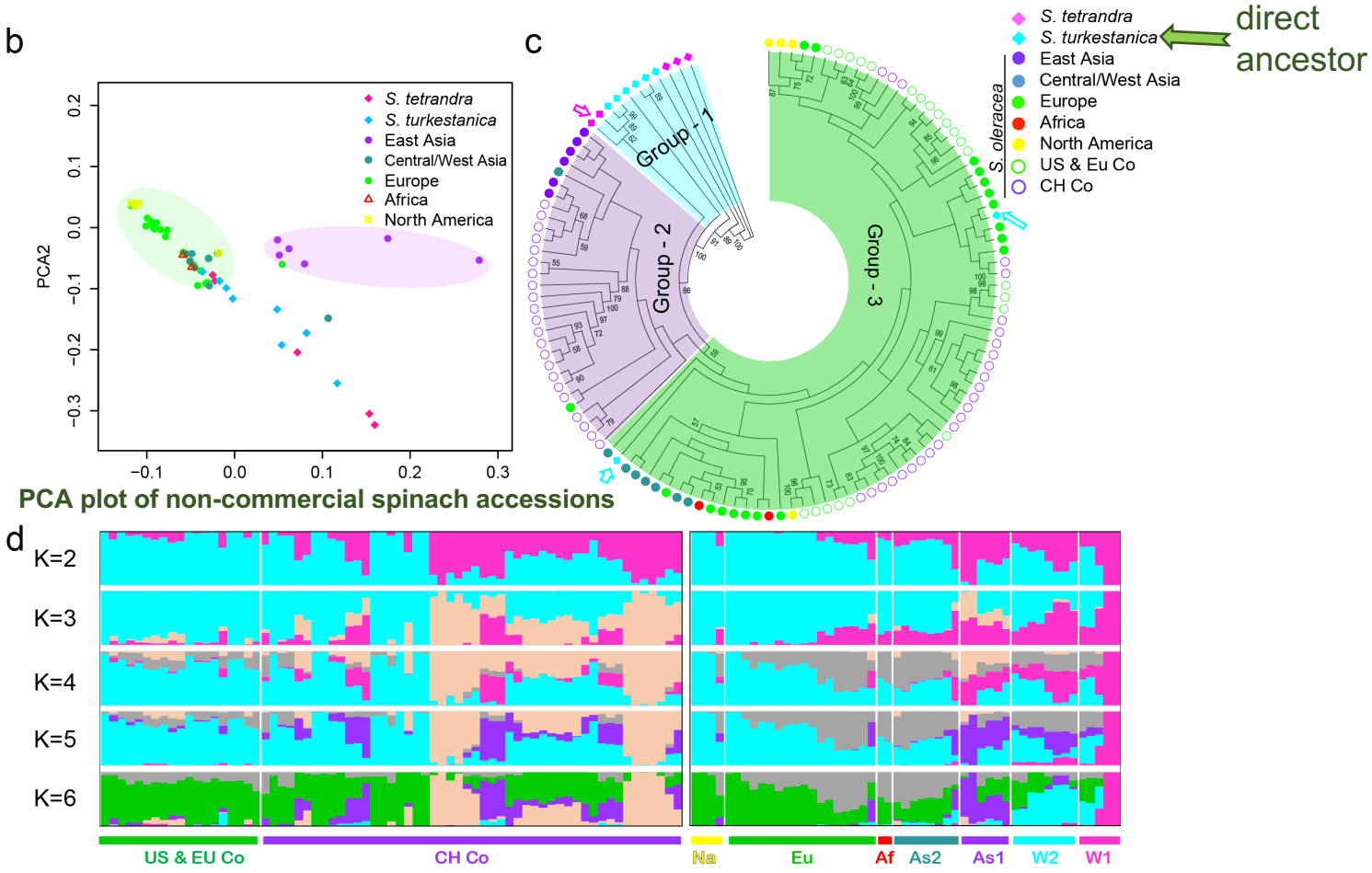
- Summary of transcriptome SNPs and small indels

Sample	Sample size	Genotype percentage	No. small indels	No. SNPs	Total
All three species	120	All	12,618	420,545	433,163
		50%	6,229	274,399	280,628
		90%	3,048	142,941	145,989
S. oleracea	107	All[a]	6,333	192,515	198,848
		50%	2,339	115,401	117,740
		90%	829	50,872	51,701
S. tetrandra (excluding Sp39 and Sp40)	3	All	2,652	117,299	119,951
		90%	1,618	88,027	89,645
S. turkestanica (excluding Sp47 and Sp48)	6	All	1,958	51,977	53,935
		50%	1,543	44,132	45,675
		90%	419	16,081	16,500

- Genetic diversity in different spinach populations

Group Description	Species	No. accessions	π per kb
All cultivars	*S. oleracea*	107	0.6687
All wild	*S. turkestanica* & *S. tetrandra*	13	4.1554
Wild accessions (excludingSp47 & Sp48)	*S. turkestanica* & *S. tetrandra*	11	4.6847
Wild *S. turkestanica accessions*	*S. turkestanica*	8	0.8217
Wild *S. turkestanica accessions* (excluding Sp47 & Sp48)	*S. turkestanica*	6	0.8323
Wild *S. tetrandra* accessions	*S. tetrandra*	5	7.2573
Wild *S. tetrandra* accessions (excluding Sp39 & Sp40)	*S. tetrandra*	3	6.4027

Population Genomics

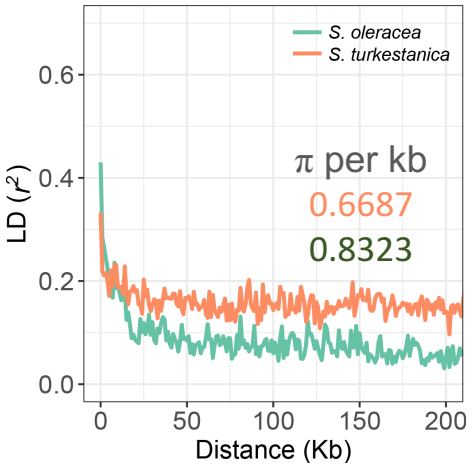


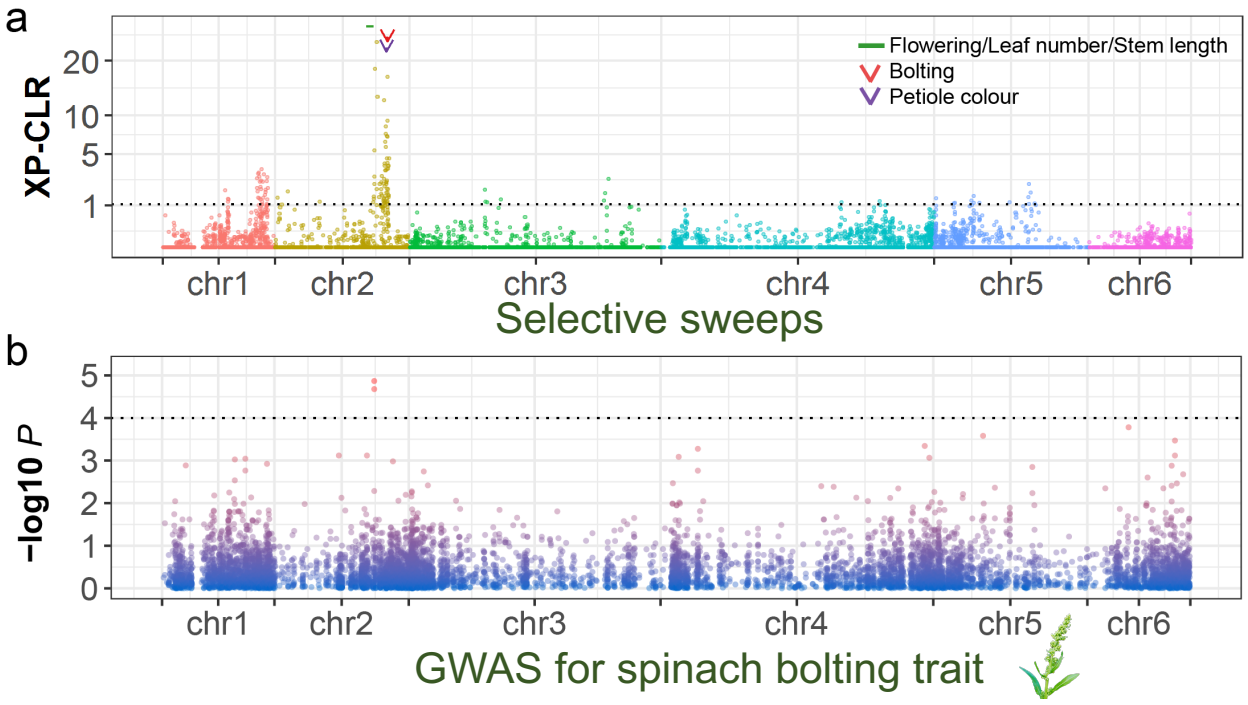
CH Co: companies in China;
US & EU Co: companies in United States and Europe.

W1: *S. tetrandra*; W2: *S. turkestanica*
As1: East Asia; As2: Central/West Asia
Eu: Europe; Af: Africa; Na: North America;

Spinach domestication

- Domestication history
  - no direct historical record or research of early domesticated traits is currently available.
  - spinach domestication has a weak bottleneck.
- Candidate domestication sweeps
  - 93 regions (~2.3 Mb), involved 261 (1.0%) protein-coding genes.



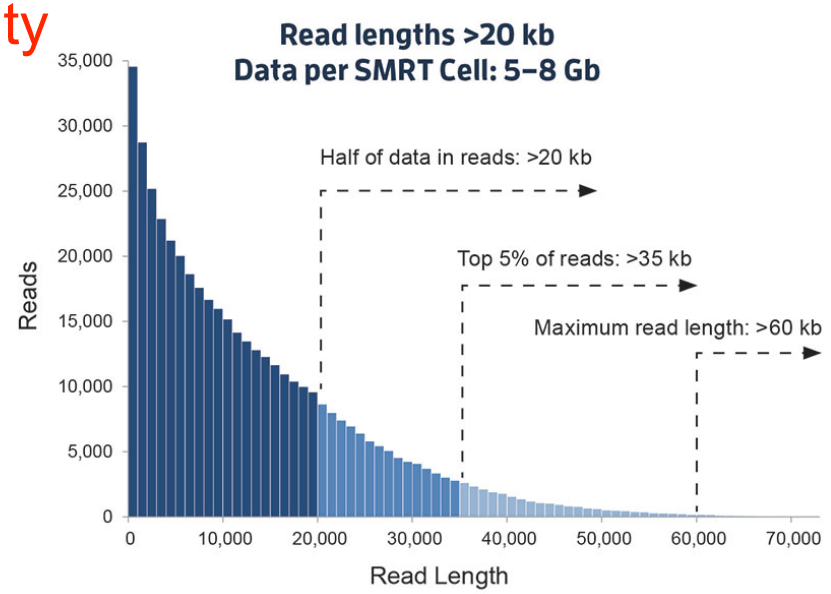


known SNP markers and QTL

Future improvement of spinach genome assembly

- Certain level of heterozygosity (sibling cross)
- Highly repetitive

PacBio + Hi-C



- Hi-C is a method using high-throughput sequencing to find the 3D architecture of whole genomes (spatial organization of chromatin)
- Quantify the number of interactions between genomic loci that are nearby in 3-D space, but could be separated by many nucleotides in the linear genome.
- Connect assembled scaffolds to pseudochromosomes

Summary

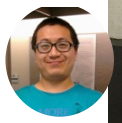
- Assembled and annotated a draft genome of spinach.
- 139 R genes in the spinach genome and five of them close to a known downy mildew resistance marker.
- No whole genome duplication in spinach. High synteny with sugar beet genome. Substantial genome rearrangements within Caryophyllales
- Small differences of genetic diversity between cultivated and wild progenitor and rapid LD decay in cultivated spinach, indicate a very weak bottleneck of spinach domestication.
- Identified 93 selective sweeps in the spinach genome (261 genes) that contain a number of QTLs and markers that are known to be associated with potential domestication traits in spinach such as bolting, flowering, leaf number and stem length.

http://spinachbase.org

Acknowledgement

BTI BOYCE THOMPSON INSTITUTE

Chen Jiao
Honghe Sun
Yi Zheng
Wenli Liu
Xuepeng Sun
John Mendieta



SHANGHAI NORMAL UNIVERSITY

Quanhua Wang
Quanxi Wang
Chenxi Xu
Xiaofeng Cai
Chenhui Ge

USDA

Beiquan Mou