Data Management for Efficient Phenotypic and Genomic Selection in Applied Breeding Programs

Rich Horsley, Ana Heilman, and Tom Walk

https://www.agronomix.com/_assets/images/CloudDevices_400.jpg

http://www.extremetech.com/wp-content/uploads/2013/03/MinION_117.jpg

https://www.bernardmarr.com/img/blog/how-big-data-is-transforming-every-business-in-every-industry.png

NDSU NORTH DAKOTA STATE UNIVERSITY

Table 1. Barley yields (bu/ac) of 2ND25276 and other barley cultivars grown in North Dakota barley breeding program yield trials, 2007-2014.

	Location								
Entry	Carrington	Fargo	Langdon/ Osnabrock	McVille	Minot	Nesson Valley	Sidney	Williston	Mean
Sta. Years	12	19	6	17	11	12	10	16	105
2ND25276	112.5	78.5	107.6	63.4	103.2	115.8	116.0	70.7	91.3
Conlon	101.5	65.2	97.6	59.5	81.9	100.3	105.8	73.1	81.7
Pinnacle	97.8	71.8	108.3	60.0	92.3	115.1	112.8	72.3	86.7
Lacey	112.3	74.8	100.8	58.6	98.6	112.1	116.4	70.2	88.4

Table 2. Malt quality comparisons of 2ND25276 and other barley cultivars grown in North Dakota yield trials, 2007-2014[†].

Entry	Barley protein (%)	Plump kernels (%)	Malt extract (%)	Wort protein (%)	S/T (%)[‡]	Diastatic power (°L)	Alpha-Amylase (20° DU)	Beta-glucan (ppm)
Sta. Yrs.	33	40	26	26	26	26	26	26
2ND25276	11.7	92.3	81.8	5.42	47.8	118	89.8	154
Conlon	13.3	93.9	80.3	5.14	39.5	138	78.0	366
Pinnacle	11.7	90.3	81.6	4.91	43.8	116	70.2	227
Lacey	13.7	85.8	79.4	5.68	42.6	189	77.5	125

[†]Data courtesy of the USDA-ARS Cereal Crops Research Unit, Madison, WI.
[‡] Soluble protein to total protein ratio.

NDSU NORTH DAKOTA STATE UNIVERSITY

Table 3. Disease reaction in the greenhouse to spot-form net blotch disease for 2ND25276 and representative cultivars entered in the Mississippi Valley Barley Nursery (MVBN), 2010-2013[†].

	Greenhouse Spot-Form Net Blotch (rating)[‡]				
	2010	2011	2012	2013	Mean
2ND25276	3.0	2.0	5.0	2.0	3.0
Morex	7.0	7.0	5.0	3.0	5.5
Robust	3.0	3.0	3.0	3.0	3.0
Legacy	3.0	3.0	3.0	3.0	3.0
Lacey	3.0	3.0	7.0	2.0	3.8
Tradition	5.0	3.0	7.0	3.0	4.5
Pinnacle	7.0	7.0	5.0	7.0	6.5

[†]Data courtesy of Agriculture and Agri-Food Canada – Cereal Research Centre, Winnipeg, Manitoba.
[‡]Reaction categories: 10=VS, 9=S, 7=MS, 5=MR-MS, 3=MR, 1=R

Table 4. Disease reaction to spot blotch disease for 2ND25276, representative cultivars and resistant and susceptible controls in field tests in 2011-2014[†].

Entry	Row type	Field Spot Blotch (percent)[‡]				
		2011	2012	2013	2014	Mean
2ND25276	2	38.3	18.3	6.3	18.3	20.3
Conlon	2	30.0	53.3	20.0	20.0	30.8
Lacey	6	18.3	8.3	6.7	6.3	9.9
Pinnacle	2	25.0	40.0	12.3	23.3	25.2
Rawson	2	35.0	16.7	9.0	16.7	19.4
Stellar-ND	6	25.0	10.7	5.0	11.3	13.0
Tradition	6	28.0	11.7	5.7	7.3	13.2
NDB112[§]	6	25.0	11.7	5.0	5.0	11.7
ND5883[¶]	6	60.0	46.4	28.4	45.0	45.0

[†]Data courtesy of Drs. Stephen Neate and Robert Brueggeman, Department of Plant Pathology, NDSU.
[‡]Numbers indicate percentage of leaf surface covered by lesions.
[§]Resistant control
[¶]Susceptible control

NDSU NORTH DAKOTA STATE UNIVERSITY

Table 5. Deoxynivalenol (DON) accumulation for 2ND25276, representative cultivars, and Fusarium head blight (FHB) resistant and susceptible controls in field tests in 2009-2013. These trials were irrigated and inoculated and typically gave higher disease and DON than experienced by farmers.

Entry	Row type	DON (ppm)[†]					
		2009	2010	2011	2012	2013	Mean
Sta. Yrs.		2	2	2	2	2	10
2ND25276	2	47.5	21.4	5.3	7.4	4.6	17.2
Robust	6	59.5	35.6	7.5	11.1	6.2	24.0
Lacey	6	55.0	24.9	10.5	15.5	5.3	22.2
Tradition	6	41.9	30.2	9.6	6.5	5.0	18.6
Stellar-ND	6	35.3	23.8	7.9	8.8	8.8	16.9
Quest[‡]	6	34.0	24.9	4.4	10.4	3.2	15.3
Conlon	2	28.8	20.1	3.1	3.1	3.0	11.6
Pinnacle	2	65.7	25.6	5.3	6.9	3.8	21.4
ND20493[§]	6	12.5	4.0	5.8	3.5	3.4	5.8

[†]Data courtesy of Dr. Paul Schwarz, NDSU Department of Plant Sciences.
[‡] FHB-resistant cultivar
[§] Resistant control

NDSU Breeding Scheme

Year	Generation	Location	Process	Number of plants or lines	Number of locations
1	Crossing	Greenhouse (fall)	Crossing block	200 crosses	1
1	F_1	Greenhouse (winter)	F_1 increase	5,000 plants	1
1	F_2	North Dakota (summer)	F_2 selection	400,000 plants	1
2	F_3	New Zealand or Puerto Rico	F_3 advancement	85,000 plants	2
2	F_4	North Dakota	Progeny rows	25,000 lines	1
3	F_5	New Zealand or Arizona	Seed increase	5,000 lines	2
3	F_6	North Dakota	Preliminary yield trials	1,200 lines	2
4	F_7	North Dakota	Intermediate yield trials	160 lines	7
5	F_8	North Dakota	Advanced yield trial	40 lines	7
6	F_9	North Dakota	Variety yield trial (VYT)	10 lines	7
7	F_{10}	North Dakota & region	VYT and AMBA Pilot Scale	8 lines	14
8	F_{11}	North Dakota & region	VYT and AMBA Pilot Scale	4-8 lines	14
9	F_{12}	North Dakota & region	VYT and AMBA Plant Scale	2 lines	20
10	F_{13}	North Dakota & region	VYT and AMBA Plant Scale	2 lines	20

How Long Does it Take to Make These Tables?

- Balanced data vs. unbalanced data.
- How do you make similar tables for all lines in the program that have been tested two or more years?
- 200 lines * 30 minutes/table = 12.5 working days
- 200 lines * 4 hours/table = 20 weeks

How NDSU Is Addressing These Data Management Needs?

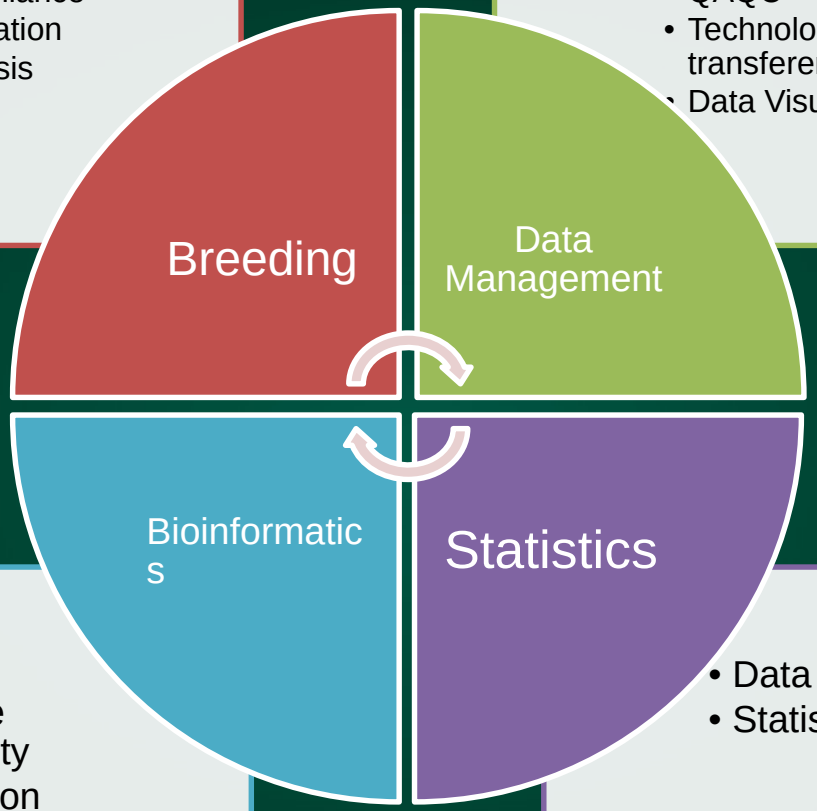
- Department identified bioinformaticists as a priority for the 2013 SBARE & Legislative process.
- Hiring bioinformaticists identified as the top priority for the AES in the 2015 SBARE & Legislative process.
- Two Breeding Pipeline Data Managers hired in summer 2016.

https://www.ag.ndsu.edu/plantsciences/news/large-data-managers-join-plant-sciences

Pipeline Database Management Team

- Data Compliance
- Standardization
- Data Analysis
- Efficiency

- QAQC
- Technology transference
- Data Visualizations



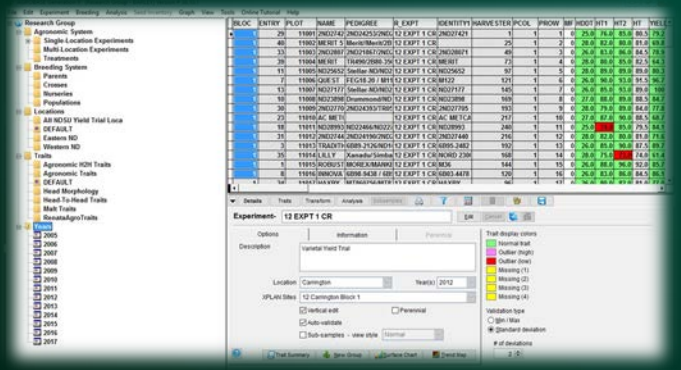
- Scripting language
- IT Security
- Automation

- Data migration
- Statistics

NDSU NORTH DAKOTA STATE UNIVERSITY

Requirements for Efficient Data Management

- Single platform for managing
  - phenotype data
  - genotype data
- Single server for housing platforms and data.
- Automating routine processes (queries and calculation of means or BLUPs)



Why use a relational database?

Flat File Database	Relational Database
Advantages	
• Easy to use, quick start	• High level of security (authentication process)
• No extra installation required	• Powerful data organization and storage
• Analytic & graphing tools built-in	• Data centralized, standardized & easily shared
	• Changes universally applied to all tables
	• Data easily queried, compiled, and analyzed
	• Intellectual property protection & compliance
Disadvantages	
• Multiple files difficult to compile & compare	• Complex set-up
• Only understood by primary user in most cases	• Changes may be difficult
• Prone to corruption	• Database needs to be well planned
• More difficult to secure	• Requires experienced management
• Data must be applied individually to each table	

Managing Phenotype Data

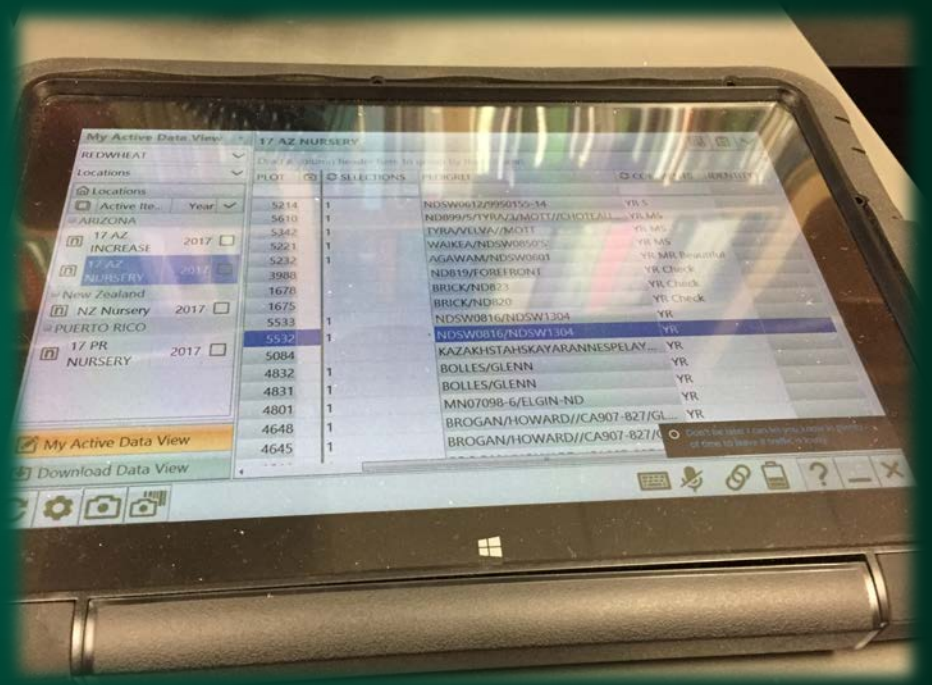
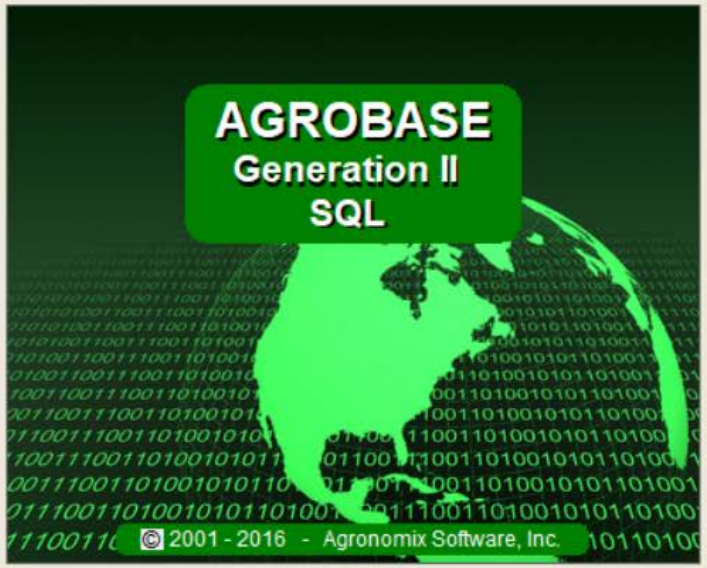


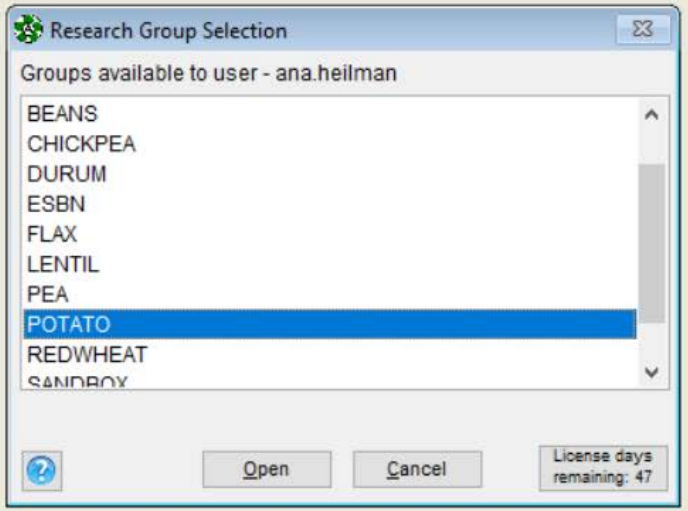
Photo courtesy: Dr. Andrew J Green

NDSU NORTH DAKOTA STATE UNIVERSITY

- Breeding system management
- Randomizations
- Data collection on tablets
- Analyses of basic experimental designs
- Stores data and means in a relational database

NDSU NORTH DAKOTA STATE UNIVERSITY


AGROBASE
Generation II
SQL
© 2001 - 2016 - Agronomix Software, Inc.

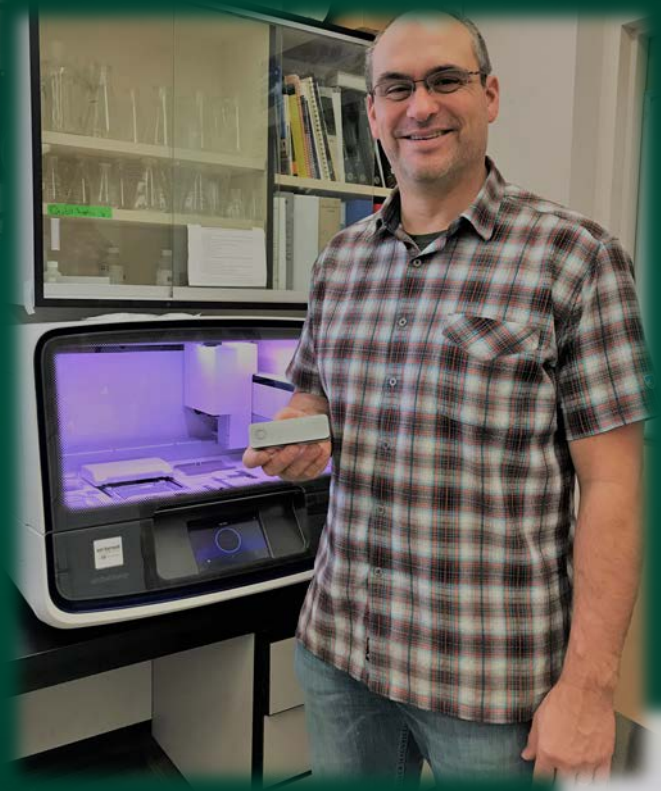

Research Group Selection
Groups available to user - ana.heilman
BEANS
CHICKPEA
DURUM
ESBN
FLAX
LENTIL
PEA
POTATO
REDWHEAT
Open
Cancel
License days remaining: 47

Data Management Work Plan at NDSU

- Educate and support breeders on use of technology platforms (e.g. Agrobase, T.3, JMP, and SAS)
- Integrate legacy data from breeding programs into DB.
- Automate data queries and calculations of:
  - Unbalanced means or BLUPS across years based on
    - **Entry**
    - **Pedigree**
    - **Parents**
  - Balanced means across years
  - Breeder requests
- Expand visualization capabilities for making decisions

Managing Genotype Data

File-DNA-Sequencers_from_Flickr_57080968.jpg



https://thumbs.dreamstime.com/z/genetic-fingerprinting-as-fingerprint-dna-emerging-out-as-medical-identification-symbol-paternity-test-64869820.jpg

https://www.extremetech.com/wp-content/uploads/2013/03/minion-usb-dna-sequencer.jpg

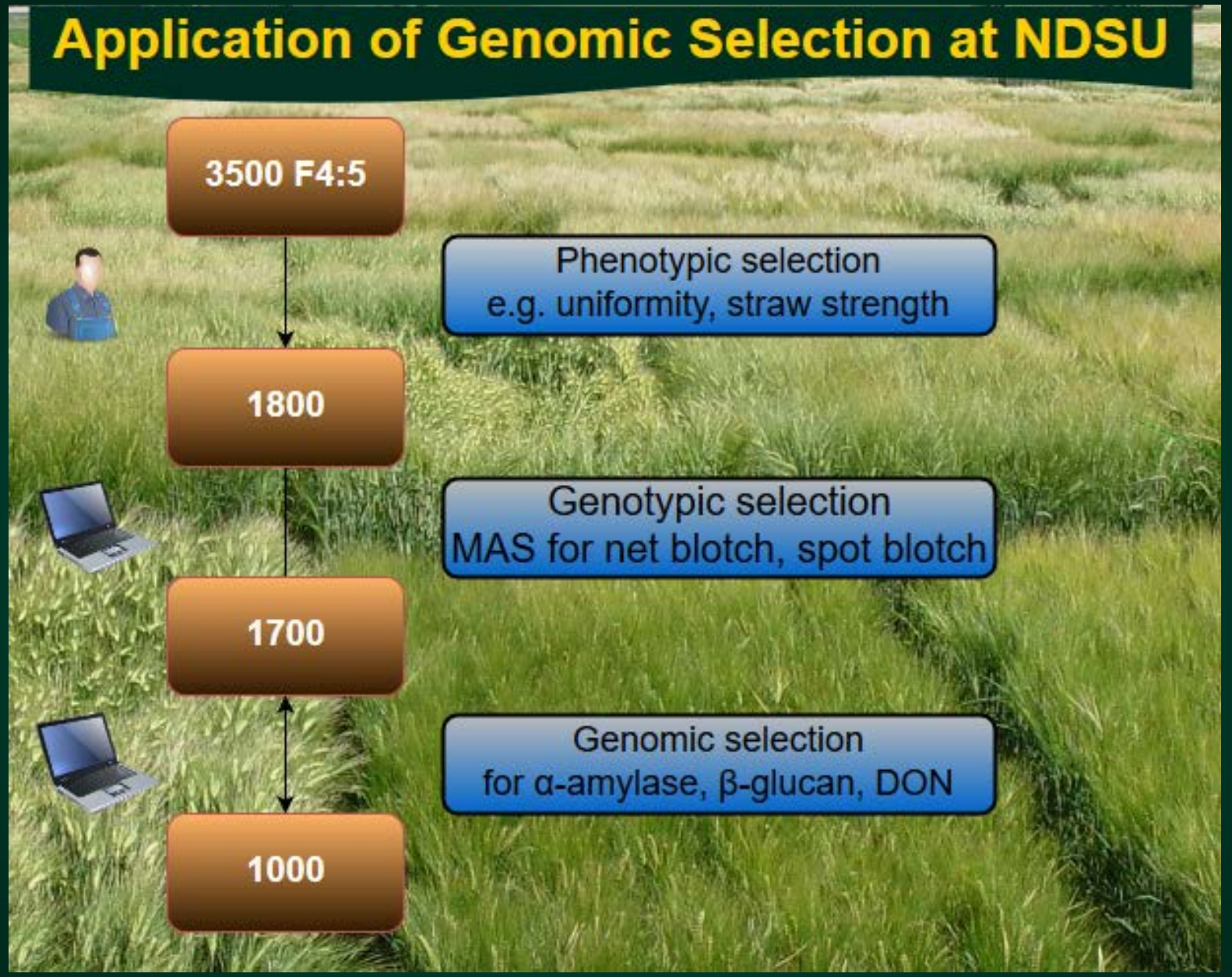
NDSU NORTH DAKOTA STATE UNIVERSITY

How do we Manage Genotype Data?

- 525 parents screened with 50,000 SNPs
  - 28.25 million data points
- 3,500 lines screened with 450 SNPs
  - 1.575 million data points
- How do you pair genotype data with phenotype data for GWAS or genomic selection?

NDSU Breeding Scheme

Year	Generation	Location	Process	Number of plants or lines	Number of locations
1	Crossing	Greenhouse (fall)	Crossing block	200 crosses	1
1	F_1	Greenhouse (winter)	F_1 increase	5,000 plants	1
1	F_2	North Dakota (summer)	F_2 selection	400,000 plants	1
2	F_3	New Zealand or Puerto Rico	F_3 advancement	85,000 plants	2
2	F_4	North Dakota	Progeny rows	25,000 lines	1
3	F_5	New Zealand or Arizona	Seed increase	5,000 lines	2
3	F_6	North Dakota	Preliminary yield trials	1,200 lines	2
4	F_7	North Dakota	Intermediate yield trials	160 lines	7
5	F_8	North Dakota	Advanced yield trial	40 lines	7
6	F_9	North Dakota	Variety yield trial (VYT)	10 lines	7
7	F_{10}	North Dakota & region	VYT and AMBA Pilot Scale	8 lines	14
8	F_{11}	North Dakota & region	VYT and AMBA Pilot Scale	4-8 lines	14
9	F_{12}	North Dakota & region	VYT and AMBA Plant Scale	2 lines	20
10	F_{13}	North Dakota & region	VYT and AMBA Plant Scale	2 lines	20



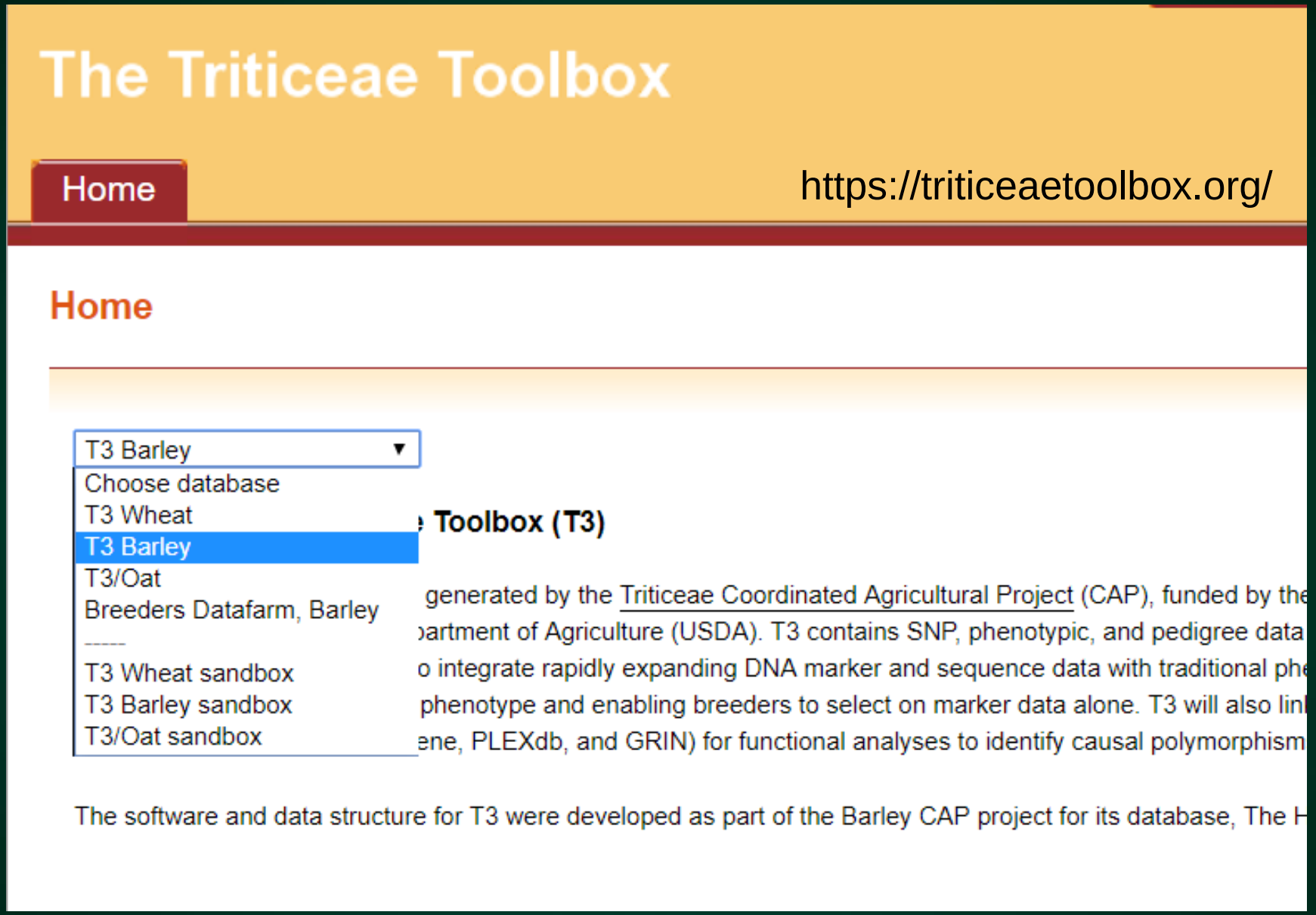
NDSU NORTH DAKOTA STATE UNIVERSITY

Favorable Phenotypes in the Arizona Winter Nursery

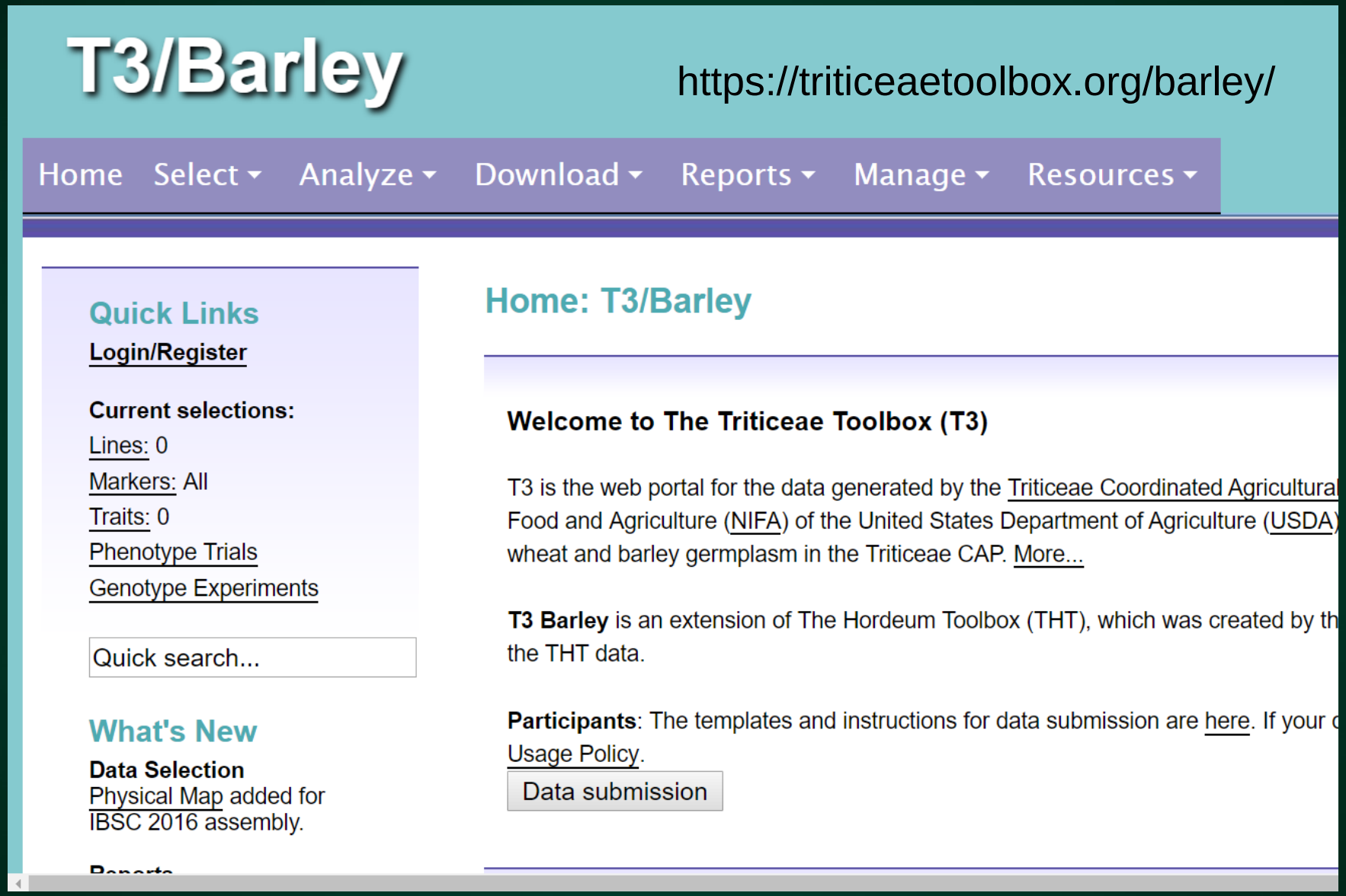
NDSU NORTH DAKOTA STATE UNIVERSITY

Weak Strawed 2-rowed Phenotypes in the Arizona Winter Nursery

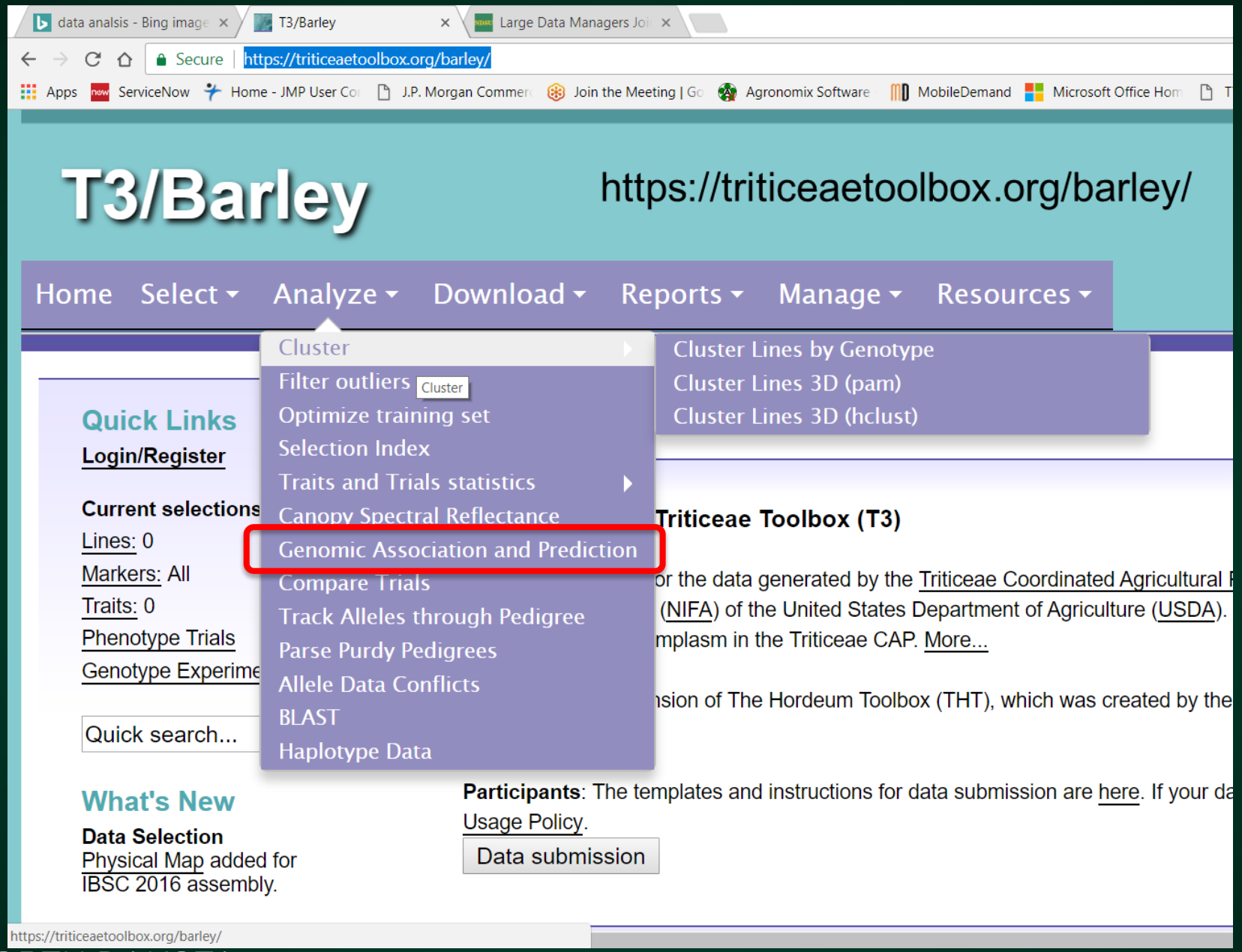
NDSU NORTH DAKOTA STATE UNIVERSITY



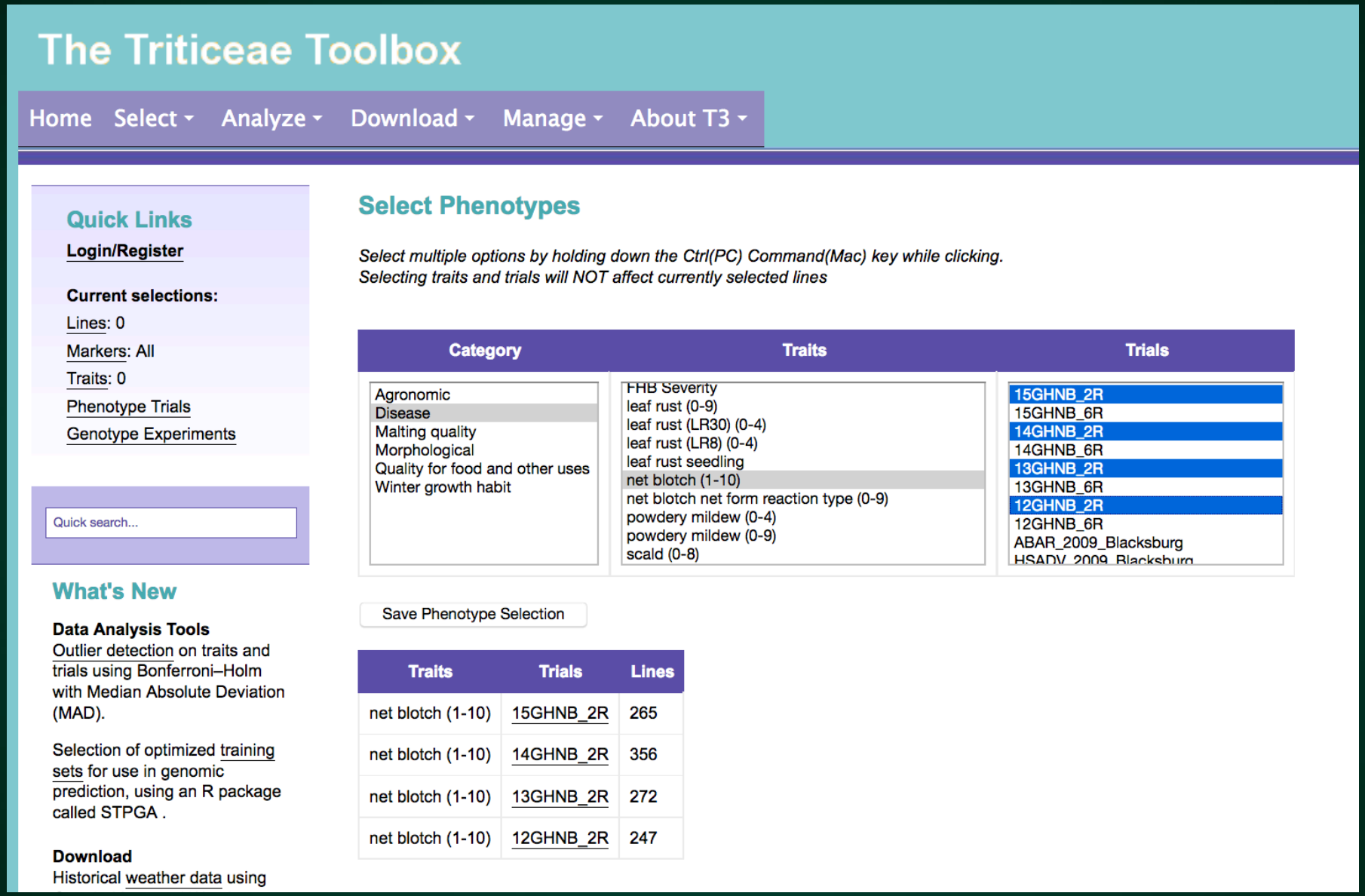
NDSU NORTH DAKOTA STATE UNIVERSITY



NDSU NORTH DAKOTA STATE UNIVERSITY



NDSU NORTH DAKOTA STATE UNIVERSITY



NDSU NORTH DAKOTA STATE UNIVERSITY

Having Data in Databases Allows for Optimization of:

- Markers to include for genomic selection.
- Lines to use in training population.
- Statistical models to use for genomic selection of specific traits.
- Predictions of progeny performance from planned crosses.

Next Steps for NDSU Data Management Team

- Determine needs for hosting T3 locally and updating system to handle more crops.
- Develop Addins in JMP for automating queries and analyses.
- Make applications available on tablet and mobile devices for NDSU breeding programs.

Any Questions?

Yield Trials at Osnabrock, ND

NDSU NORTH DAKOTA STATE UNIVERSITY