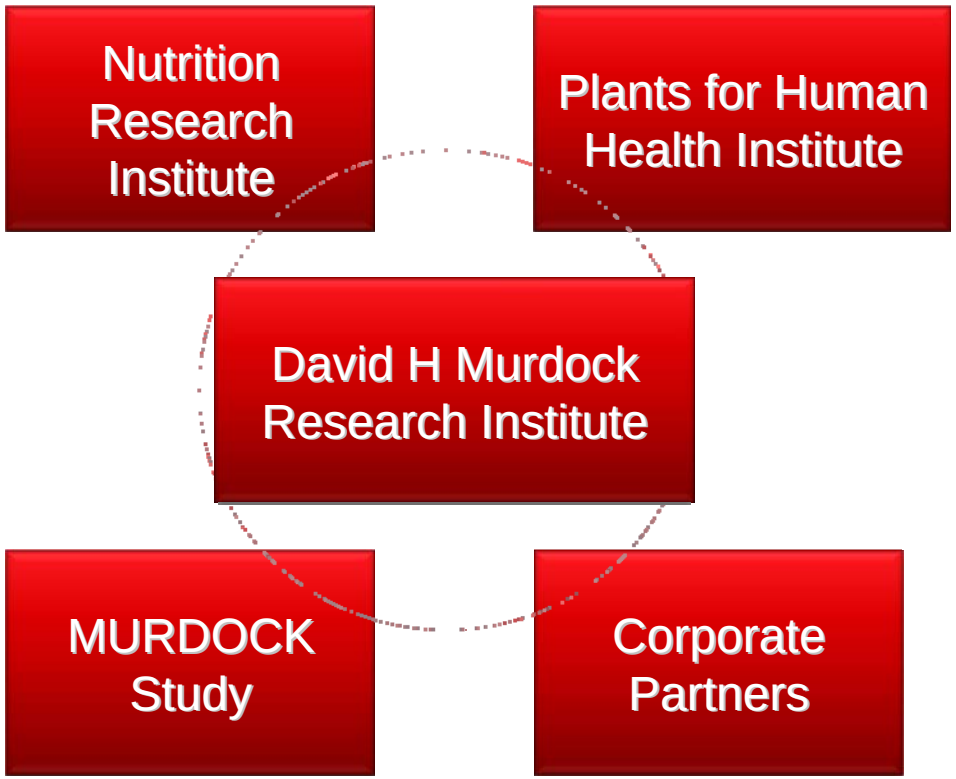
Blueberry Genomics: Practical Applications from New Technologies

Allan Brown, PhD
Plants for Human Health Institute
NC State University, Kannapolis NC

North Carolina Research Campus

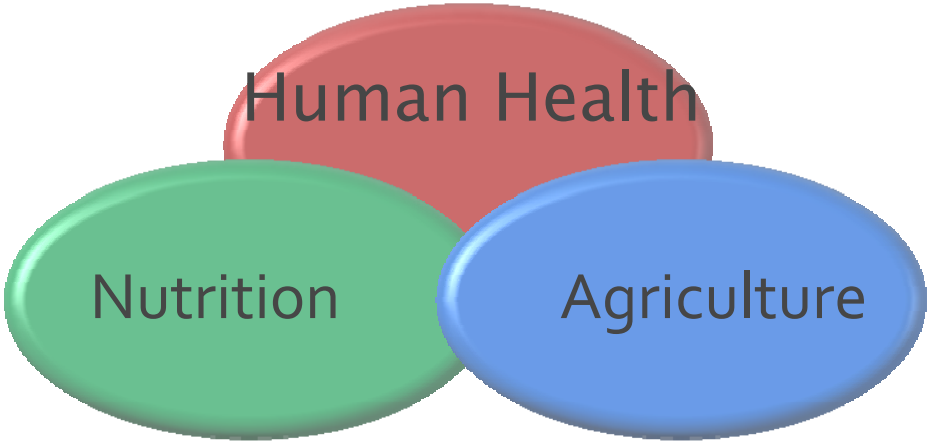


North Carolina State University • UNC Chapel Hill • UNC Charlotte • UNC Greensboro • Duke University • North Carolina Central University • North Carolina A&T University • Appalachian State University • Rowan Cabarrus Community College • LabCorp • Anatomics • Carolinas HealthCare System • Dole • IMAF • Lovelace • redhat • Sensory Spectrum • Zeiss

A broad range of integrated multidisciplinary technologies are available at the DHMRI

. . . all under one roof!

NCRC & the DHMRI



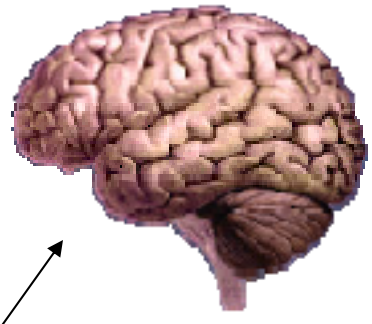
- Our mission is to be the leader at the intersection of human health, nutrition and agriculture using cutting edge science and technology platforms
- Develop unique partnerships via close collaborations to ensure the NCRC becomes a world class scientific community
- Provide new insights and novel interpretation through impactful science

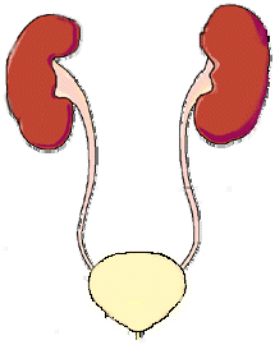
Health benefits

- Flavonoids, Anthocyanins, Proanthocyandins
- High antioxidant capacity
- Apoptosis, anti-inflamation, modulation of the MAPK signaling pathway and enhanced induction of xenobiotic detoxification enzymes

Berries and Human Health

American Cancer Society®





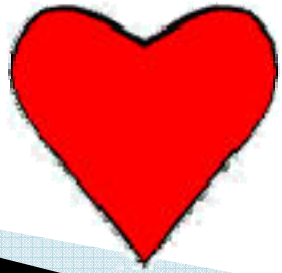


Table 8--Cultivated blueberries: Commercial acreage, yield per acre, production, and season-average grower price in the United States, 1980-2008

State and year	Acreage harvested	Yield per acre	Utilized production	Utilization		Grower price			Value of utilized production
				Fresh	Processed	Fresh	Processed	All	
	Acres	*Pounds*		*-- 1,000 pounds*			*-- Dollars/pound --*		*1,000 dollars*
1980	21,850	NA	81,063	43,183	37,885	NA	NA	NA	41,361
1981	22,180	NA	95,250	44,516	50,734	NA	NA	NA	55,161
1982	12,120	NA	85,770	43,430	42,340	NA	NA	NA	65,306
1983	25,400	NA	89,698	40,355	49,343	NA	NA	NA	61,501
1984	12,650	NA	95,376	57,084	37,292	NA	NA	NA	50,824
1985	12,500	NA	102,600	56,778	45,822	NA	NA	NA	59,009
1986	28,600	NA	111,417	53,817	57,600	NA	NA	NA	68,777
1987	13,000	NA	110,788	52,411	58,677	NA	NA	NA	77,064
1988	13,600	NA	100,502	45,904	54,214	NA	NA	NA	94,858
1989	30,420	NA	127,620	56,885	70,735	NA	NA	NA	86,842
1990	13,100	NA	103,845	51,730	52,115	NA	NA	NA	66,546
1991	13,550	NA	114,766	50,472	64,294	NA	NA	NA	88,155
1992	33,650	3,310	111,320	45,502	65,818	1.10	0.67	0.85	94,097
1993	36,500	4,600	167,748	69,545	98,203	0.88	0.33	0.56	93,254
1994	37,100	3,680	136,460	68,040	68,420	0.90	0.43	0.66	90,673
1995	38,040	4,180	159,000	74,760	84,240	0.90	0.40	0.64	101,279
1996	37,750	3,320	125,380	62,380	63,000	1.06	0.76	0.91	113,780
1997	38,670	4,310	166,620	69,300	97,320	1.10	0.64	0.83	138,490
1998	39,000	3,790	147,880	75,140	72,740	0.97	0.48	0.73	107,483
1999	39,630	4,400	174,260	77,520	96,740	1.16	0.66	0.88	153,978
2000	40,820	4,480	182,890	79,080	103,810	1.29	0.73	0.97	177,804
2001	40,430	4,670	188,750	88,290	100,460	1.26	0.53	0.87	164,059
2002	41,850	4,510	188,650	100,490	88,160	1.41	0.60	1.03	194,566
2003	42,070	4,470	187,900	103,620	84,280	1.49	0.78	1.17	220,649
2004	44,850	5,070	227,610	124,590	103,020	1.55	0.81	1.21	276,011
2005	48,980	4,860	238,210	123,140	115,070	1.93	0.91	1.44	342,347
2006	54,440	5,210	283,650	146,860	136,790	2.11	1.39	1.76	500,052
2007	53,420	5,370	286,780	149,830	136,950	2.14	1.54	1.85	531,075
2008 1/	60,180	5,790	348,660	191,860	156,800	2.11	0.86	1.54	538,674

Table 2--U.S. blueberry production and utilization (cultivated and wild), selected States, 1980-2008

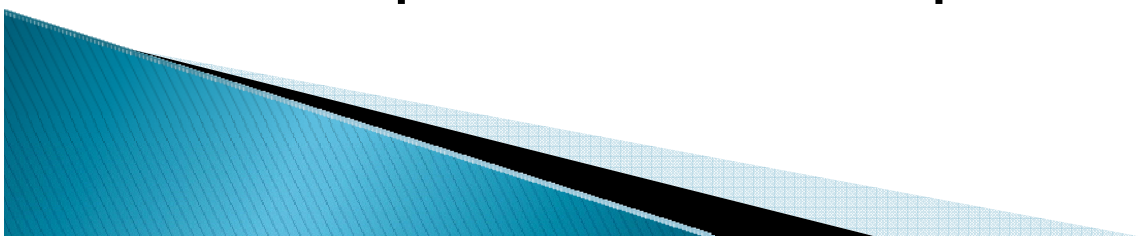
Item/State	2000	2001	2002	2003	2004	2005	2006	2007	2008 P
				1,000 pounds					
Total utilized production:									
Maine 1/	110,990	75,200	62,400	80,400	46,000	60,150	74,600	77,250	89,950
Michigan 2/	62,000	70,000	64,000	62,000	80,000	66,000	90,000	93,000	110,000
New Jersey	34,000	37,000	42,000	40,000	39,000	45,000	52,000	54,000	59,000
North Carolina	17,500	13,500	15,500	22,500	22,900	26,000	26,600	16,200	28,500
Oregon	29,000	28,500	26,500	23,900	34,000	34,500	35,600	45,000	43,100
Washington	12,410	15,000	13,650	13,200	18,000	19,600	19,000	29,600	32,000
Alabama	450	530	430	450	610	560	350	410	360
Arkansas	1,330	1,120	1,770	1,550	1,800	1,350	1,600	70	800
Florida	2,800	3,100	2,900	3,500	5,600	5,200	7,000	7,800	9,800
Georgia	19,000	17,000	17,000	17,000	21,000	26,000	31,500	11,000	41,000
Indiana	2,500	1,500	3,000	1,800	3,000	3,500	3,400	1,400	3,800
New York	1,900	1,500	1,900	2,000	1,700	1,400	2,000	2,300	2,300
California 4/	NA	NA	NA	NA	NA	9,100	10,000	16,500	14,000
Mississippi 5/	NA	NA	NA	NA	NA	NA	4,600	9,500	4,000
Total	293,880	263,950	251,050	268,300	273,610	298,360	358,250	364,030	438,610
Fresh use:									
Maine 1/	420	350	400	400	300	350	400	450	550
Michigan 2/	19,000	21,000	22,000	24,000	36,000	25,000	29,000	30,000	40,000
New Jersey	24,000	29,000	37,000	33,000	33,000	33,000	40,000	41,000	46,000
North Carolina	10,500	10,500	11,300	16,100	16,400	16,100	18,600	11,000	20,000
Oregon 3/	10,000	10,900	11,000	10,400	13,400	13,800	13,900	16,800	19,400
Washington	2,300	4,200	2,850	3,400	5,000	3,900	4,500	11,200	12,000
Alabama	450	530	430	450	610	560	350	410	360
Arkansas	1,330	1,120	1,770	1,550	1,800	1,350	1,600	70	800
Florida	2,300	2,800	2,900	3,500	5,600	5,200	7,000	7,800	9,800
Georgia	6,000	6,000	8,000	8,000	10,000	12,000	16,000	9,000	24,000
Indiana	1,500	900	1,500	1,300	1,500	2,000	1,900	800	2,200
New York	1,800	1,450	1,800	1,950	1,400	1,350	1,950	2,250	2,200
California 4/	NA	NA	NA	NA	NA	9,100	10,000	14,500	12,500
Mississippi 5/	NA	NA	NA	NA	NA	NA	2,600	5,000	2,600
Total	79,600	88,290	100,890	104,020	124,890	123,490	147,260	150,280	192,410
Processed:									
Maine 1/	110,570	74,850	62,000	80,000	45,700	59,800	74,200	76,800	89,400
Michigan 2/	43,000	49,000	42,000	38,000	44,000	41,000	61,000	63,000	70,000
New Jersey	10,000	8,000	5,000	7,000	6,000	12,000	12,000	13,000	13,000
North Carolina	7,000	3,000	4,200	6,400	6,500	9,900	8,000	5,200	8,500
Oregon 3/	19,000	17,600	15,500	13,500	20,600	20,700	21,700	28,200	23,700
Washington	10,110	10,800	10,800	9,800	13,000	15,700	14,500	18,400	20,000
Alabama	6/	6/	6/	6/	6/	6/	6/	6/	6/
Arkansas	6/	6/	6/	6/	6/	6/	6/	6/	6/
Florida	[illegible]	300	6/	6/	6/	6/	6/	6/	6/
Georgia	13,000	[illegible]	9,000	9,000	11,000	14,000	15,500	2,000	17,000
Indiana	1,000	[illegible]	1,500	500	1,500	1,500	1,500	600	1,600
New York	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

However…

- Considerable phytochemical variation among varieties
- Variation at different locations
- Screening for phytochemicals expensive and labor intensive
  - Total phenols or ORAC
  - Individual compounds
- Difficulties associated with breeding
  - 3 years until plant bears fruit
  - Inbreeding depression

Improvement of phytochemical profiles in blueberry

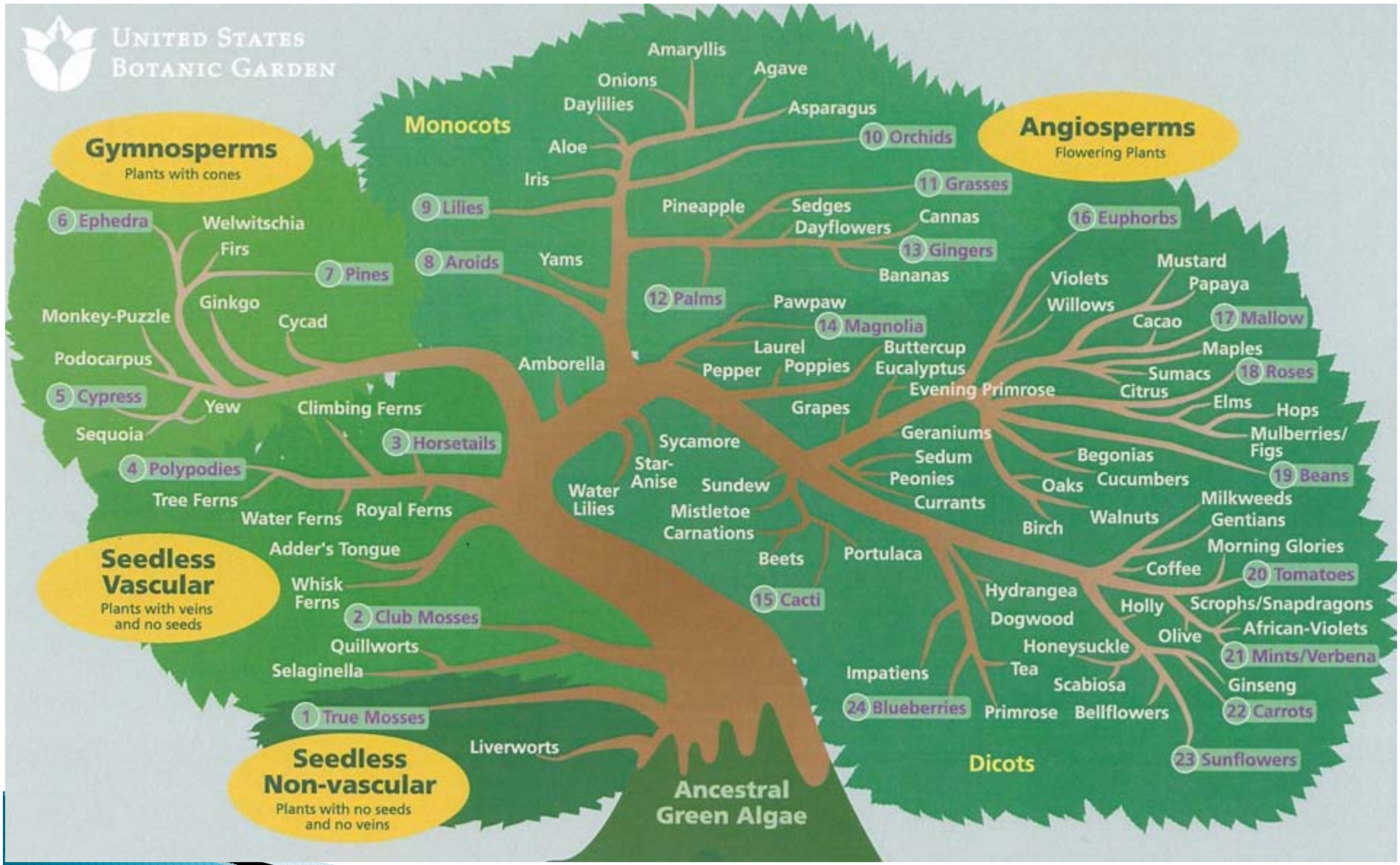
- Creation of a draft genomic sequence
- Annotation of draft sequence
- Development of high throughput analysis of phytochemicals
  - Phenolics/Flavonoids/Anthocyanins/proanthocyanin
  - Carotenoids/tocopherols/ascorbate
- Development of markers associated with candidate genes (phytochemicals)
- Association mapping of genes associated with qualitative and quantitative variability



Blueberry working group

- Allan Brown (NCSU/PHHI)
- Mary Ann Lila (NCSU/PHHI)
  - Food scientist/phytochemical analysis
- Jim Ballington (NCSU)
  - Blueberry breeding
- Ann Loraine (UNCC)
  - Bioinformatics
- Schylur Korban (Illinois)
  - BAC Library construction
- Todd Michael (Rutgers)
- DHMRI
  - Mark Burke, Steve Colman, Simon Gregory, Yankai Jia


JUN 15 2009

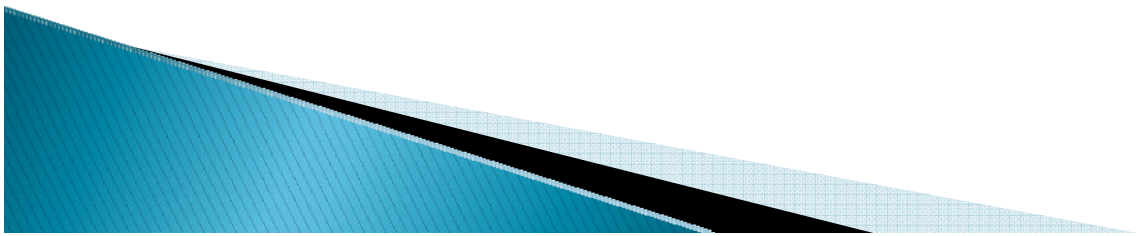

United States Botanic Garden
Gymnosperms
Plants with cones
Monocots
Angiosperms
Flowering Plants
Seedless Vascular
Plants with veins and no seeds
Seedless Non-vascular
Plants with no seeds and no veins
Dicots
Ancestral Green Algae
Amaryllis
Agave
Onions
Daylilies
Asparagus
10 Orchids
Aloe
Iris
11 Grasses
9 Lilies
Pineapple
Sedges
Dayflowers
Cannas
16 Euphorbs
6 Ephedra
Welwitschia
Firs
8 Aroids
Yams
13 Gingers
Bananas
Mustard
Papaya
7 Pines
Violets
Willows
12 Palms
Pawpaw
Ginkgo
Monkey-Puzzle
Cycad
14 Magnolia
Cacao
17 Mallow
Laurel
Buttercup
Maples
Podocarpus
Amborella
Pepper
Poppies
Eucalyptus
Sumacs
18 Roses
5 Cypress
Evening Primrose
Citrus
Elms
Hops
Yew
Climbing Ferns
Grapes
Sequoia
Geraniums
Mulberries/ Figs
3 Horsetails
Sycamore
Sedum
Begonias
4 Polypodies
Star-Anise
Peonies
Oaks
Cucumbers
19 Beans
Water Lilies
Sundew
Currants
Milkweeds
Tree Ferns
Mistletoe
Walnuts
Gentians
Water Ferns
Royal Ferns
Carnations
Birch
Morning Glories
Adder's Tongue
Beets
Portulaca
Coffee
20 Tomatoes
Whisk Ferns
15 Cacti
Hydrangea
Holly
Scrophs/Snapdragons
2 Club Mosses
Dogwood
Olive
African-Violets
Quillworts
Honeysuckle
21 Mints/Verbena
Selaginella
Impatiens
Tea
Scabiosa
Ginseng
1 True Mosses
24 Blueberries
Primrose
Bellflowers
22 Carrots
Liverworts
23 Sunflowers

Blueberry

- Family: Ericaceae
- Genus: *Vaccinum* (section:Cyanococcus) (cranberry is in section Oxycoccos)
- Multiple species
  - Highbush, lowbush, rabbiteye
- Diploid (2n=2x=24), tetraploid (2n=4x=48) and hexaploid (2n=6x=72)
- Significant inbreeding depression even in "self-compatible cultivars".
- Clonally propagated

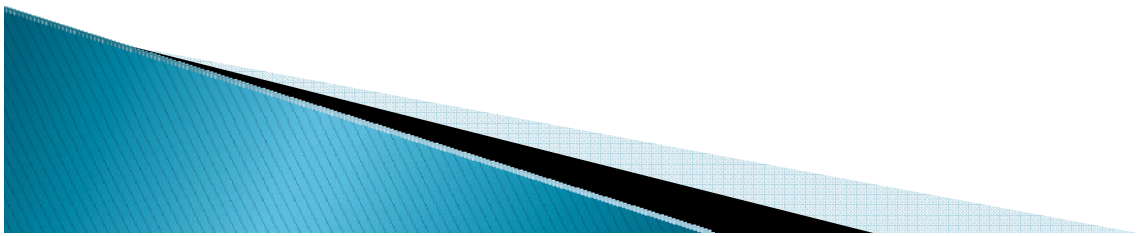
Challenges:

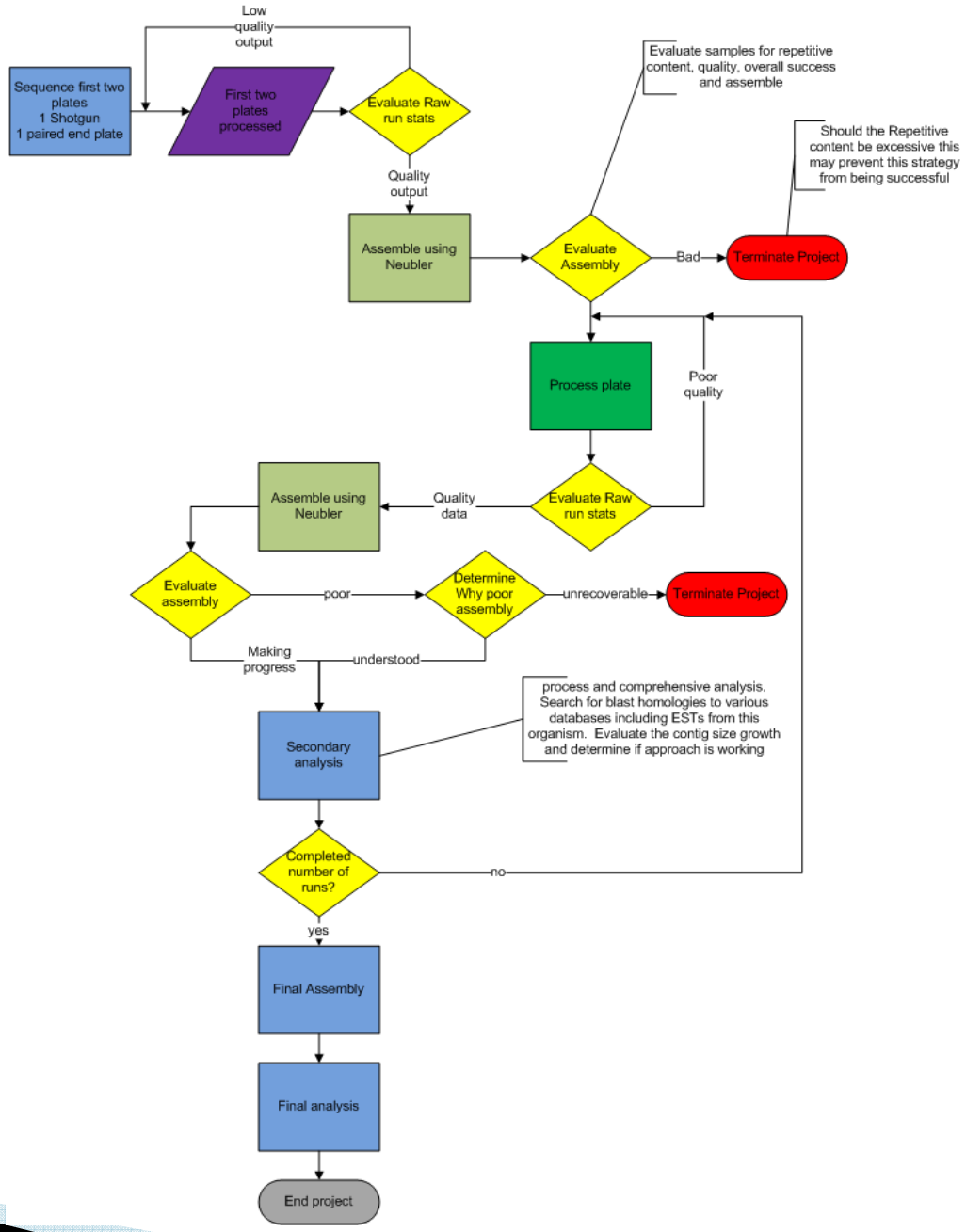
- Relatively large genome (~500 mb)
- Heterozygosity >20%
- Not closely related to model organisms
- Clonally propagated, may have large inversions/insertions/deletions
- Limited amount of sequence information available
  - ~5000 ESTs publically available
  - ~20,000 ESTs Hortresearch NZ



Strategey

- Create shotgun, 3 and 20 kb paired end libraries from‘W85-20’ (diploid Vaccinium corymbosum)
- “Fill in” scaffolds with Illumina shotgun sequences
- Create BAC library from ‘W85-20’
- Use to clarify assembly ambiguities.
- Collaborate with new cranberry sequencing project at Rutgers

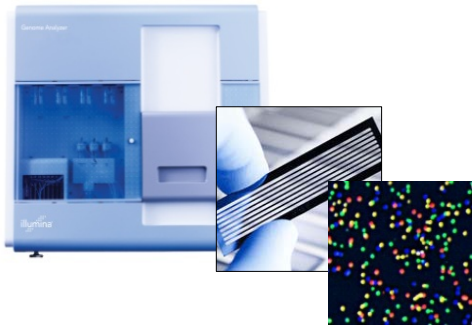



Low quality output
Sequence first two plates
1 Shotgun
1 paired end plate
First two plates processed
Evaluate Raw run stats
Quality output
Evaluate samples for repetitive content, quality, overall success and assemble
Should the Repetitive content be excessive this may prevent this strategy from being successful
Assemble using Neubler
Evaluate Assembly
Bad
Terminate Project
Process plate
Poor quality
Assemble using Neubler
Quality data
Evaluate Raw run stats
Evaluate assembly
poor
Determine Why poor assembly
unrecoverable
Terminate Project
Making progress
understood
process and comprehensive analysis. Search for blast homologies to various databases including ESTs from this organism. Evaluate the contig size growth and determine if approach is working
Secondary analysis
Completed number of runs?
no
yes
Final Assembly
Final analysis
End project

Platforms – Sequencing

Genome Analyzer from Illumina

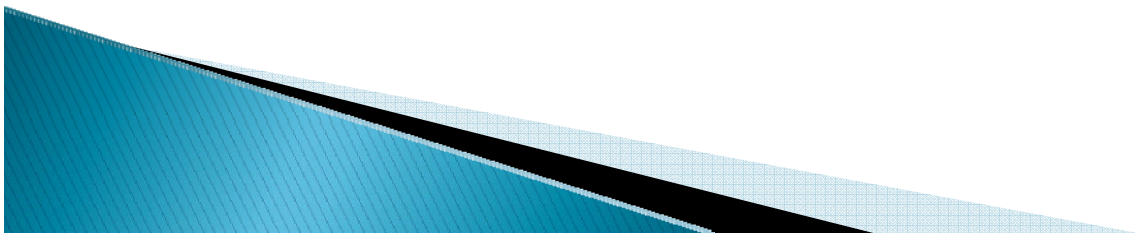
The Illumina Genome Analyzer II (GAII) utilizes a “sequencing by synthesis” approach with read lengths of **36-75**bp. This instrument is currently the industry leader in terms of cost per sequenced base and sequence reads can be obtained using a shotgun approach, from paired ends and from larger insert mate-pairs. Capacity to generate **3-4Gb** of sequence data in **5/6 days**.



Genome Sequencer FLX from Roche

The Genome Sequencer FLX System™ applications include whole genome sequencing, transcriptome analysis, gene regulation studies, and amplicon sequencing. It generates more than **400Mb** per **7.5-hou**r instrument run and achieves single-read accuracies greater than 99.5% over **400 bases per read**, and obtains consensus accuracies of greater than 99.99%. Sequence can be obtained through shotgun, paired end and mate-pair sequence reads

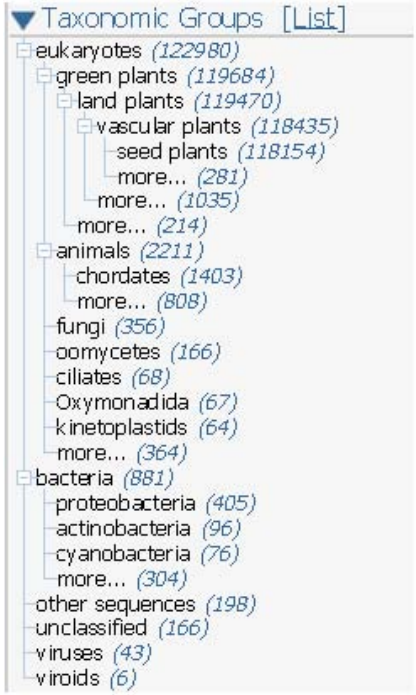
Number of Large Contigs	110,093
Large Contig Bases	87,234,876
Large Contig Average Depth	3.6
Large Contig Average Contig Size	792
Large Contig N50 Size	795
Large Contig Q40 Plus	86.68%
Number of Scaffold	109,025
Scaffold Bases	88,792,236
Scaffold Average Contig Size	814
Scaffold N50 Size	806
Number of Contigs	346,413
Total Contig Bases	158,724,623

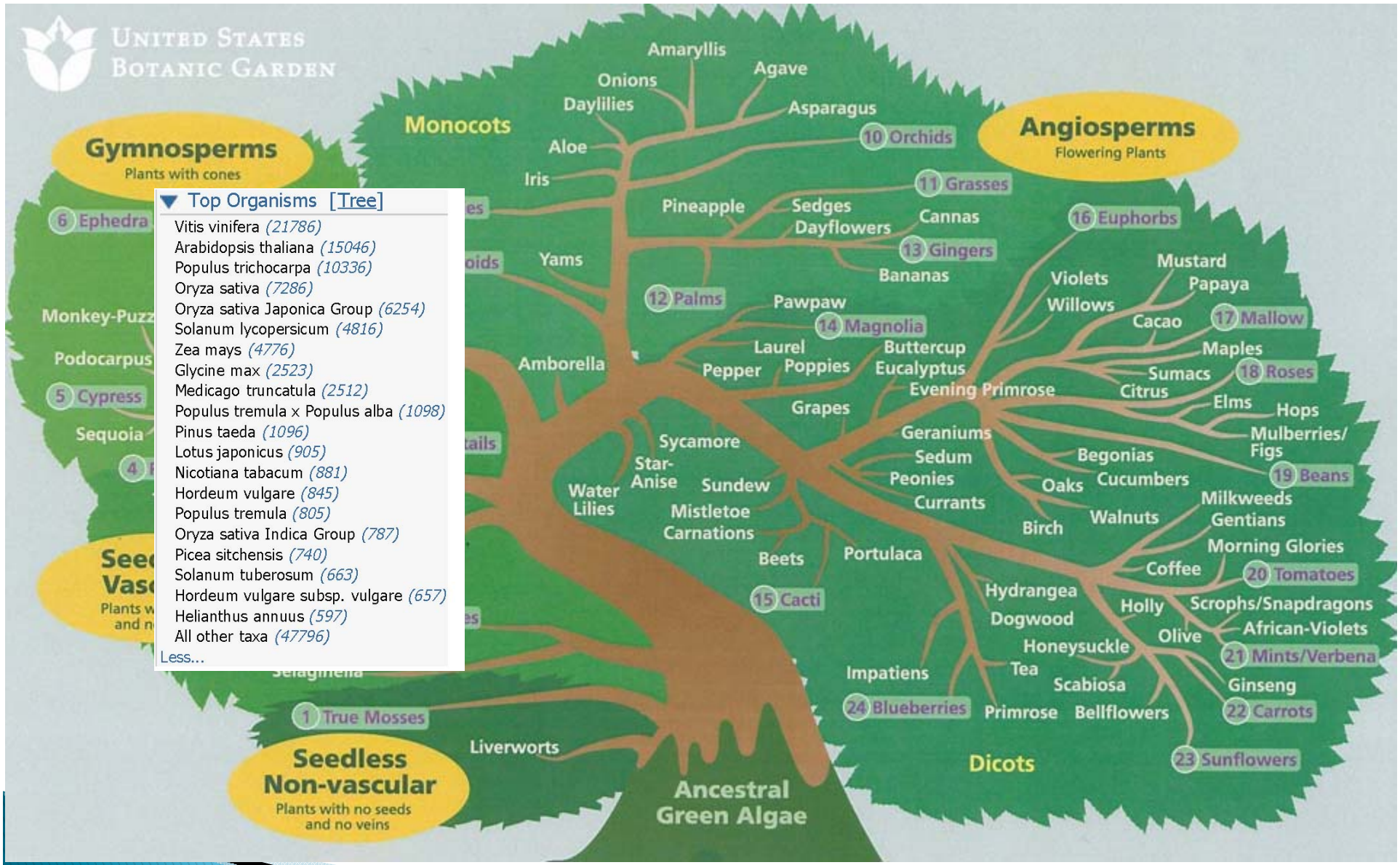


type	sub type	number of elements*	length occupied (bp)	percentage of sequence (%)
SINEs:		0	0	0
	ALUs	0	0	0
	MIRs	0	0	0
LINEs:		0	0	0
	LINE1	0	0	0
	LINE2	0	0	0
	L3/CR1	0	0	0
LTR	elements:	0	0	0
	MaLRs	0	0	0
	ERVL	0	0	0
	ERV_classI	0	0	0
	ERV_classII	0	0	0
DNA	elements:	0	0	0
	MER1_type	0	0	0
	MER2_type	0	0	0
Unclassified:		298,560,958	298,560,958	39.2
Total interspersed repeats:			298,560,958	39.2
Small RNA		0	0	0
Satellites:		1	201	0
Simple repeats:		67,191	2,523,730	0.33
Low complexity:		59,382	2,250,984	0.3

- 38% repetitive content
- RepeatMasker–5%
- RepeatScout – 38%
  - De novo predictaion

Type	number	Total Hits	Contigs with hits	% contigs with hit
Genomic Contigs (>1000 bp)	97,108	-------	-------	-------
Scaffolds (paired-end)	33,277	-------	-------	-------
Total Dataset	130,385	431762	124336	95.36%




United States Botanic Garden
Gymnosperms
Plants with cones
Angiosperms
Flowering Plants
Monocots
Dicots
Seedless Non-vascular
Plants with no seeds and no veins
Ancestral Green Algae
Top Organisms [Tree]
Vitis vinifera (21786)
Arabidopsis thaliana (15046)
Populus trichocarpa (10336)
Oryza sativa (7286)
Oryza sativa Japonica Group (6254)
Solanum lycopersicum (4816)
Zea mays (4776)
Glycine max (2523)
Medicago truncatula (2512)
Populus tremula x Populus alba (1098)
Pinus taeda (1096)
Lotus japonicus (905)
Nicotiana tabacum (881)
Hordeum vulgare (845)
Populus tremula (805)
Oryza sativa Indica Group (787)
Picea sitchensis (740)
Solanum tuberosum (663)
Hordeum vulgare subsp. vulgare (657)
Helianthus annuus (597)
All other taxa (47796)
Less...
Amaryllis
Agave
Onions
Daylilies
Asparagus
Aloe
10 Orchids
Iris
11 Grasses
Pineapple
Sedges
Dayflowers
Cannas
16 Euphorbs
13 Gingers
Yams
Bananas
Mustard
Papaya
Violets
Willows
12 Palms
Pawpaw
14 Magnolia
Cacao
17 Mallow
Maples
Laurel
Buttercup
Amborella
Pepper
Poppies
Eucalyptus
Sumacs
18 Roses
Evening Primrose
Citrus
Elms
Hops
Grapes
Geraniums
Mulberries/ Figs
Sycamore
Sedum
Begonias
Star-Anise
Peonies
Oaks
Cucumbers
19 Beans
Sundew
Currants
Milkweeds
Water Lilies
Mistletoe
Walnuts
Gentians
Carnations
Birch
Morning Glories
Beets
Portulaca
Coffee
20 Tomatoes
15 Cacti
Hydrangea
Holly
Scrophs/Snapdragons
Dogwood
African-Violets
Olive
Honeysuckle
21 Mints/Verbena
Tea
Impatiens
Scabiosa
Ginseng
24 Blueberries
Primrose
Bellflowers
22 Carrots
23 Sunflowers
1 True Mosses
Liverworts
6 Ephedra
Monkey-Puzz
Podocarpus
5 Cypress
Sequoia

Data Type	number	Unique BlastN matches (e-5)	Unique BlastN matches (e-10)
Blueberry EST "Unigenes"	3,686	-------	-------
Genomic Contigs	97,108	1,975	1,875
Paired-End Contigs (Scaffolds)	33,277	558	520
Total Matches	-------	2,533	2,395

5,135 *Vaccinium corymbosum* EST sequences in genbank

These EST files downloaded and validated in a CAP3

The resulting contigs and singlets (a.k.a. unigene set) were used as queries in TimeLogic accelerated BLAST searches against the blueberry contigs, scaffolds, and singlet datasets.

69% have been identified at e-5 (65% at e-10) using BLASTN homology searching within the assembled contigs and scaffolds.