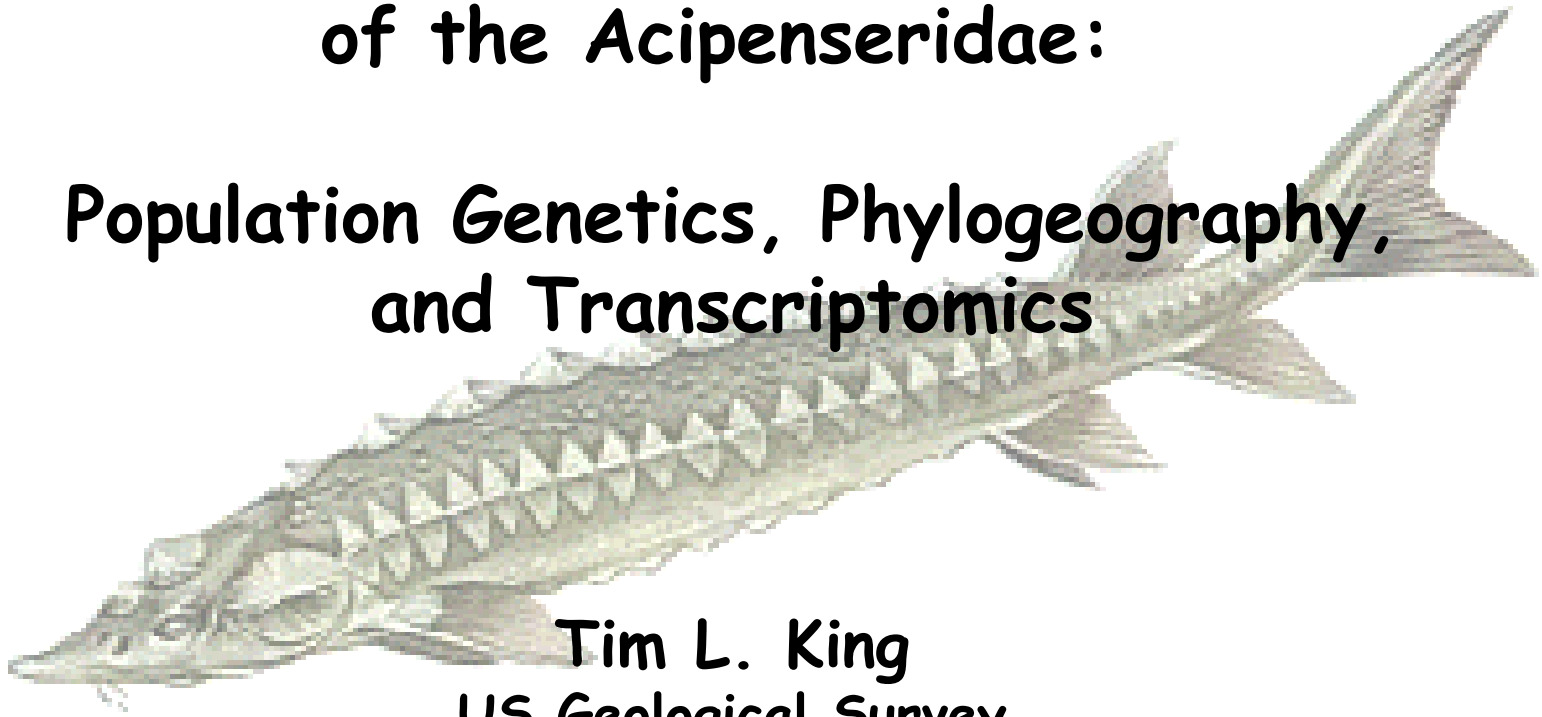
Conservation Genetics and Genomics of the Acipenseridae:

Population Genetics, Phylogeography, and Transcriptomics



Tim L. King
US Geological Survey
Leetown Science Center
Aquatic Ecology Branch

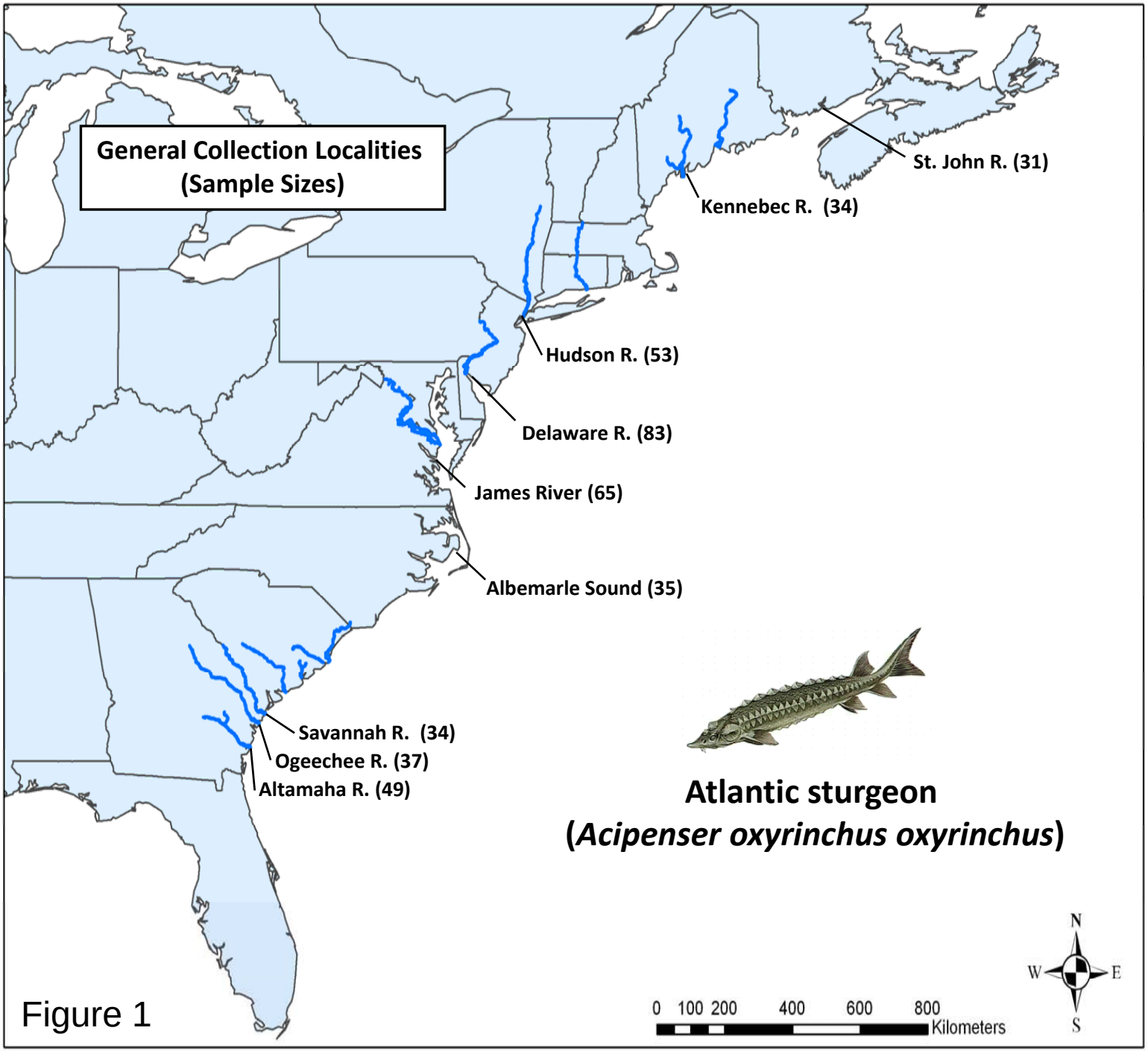

General Collection Localities
(Sample Sizes)
St. John R. (31)
Kennebec R. (34)
Hudson R. (53)
Delaware R. (83)
James River (65)
Albemarle Sound (35)
Savannah R. (34)
Ogeechee R. (37)
Altamaha R. (49)
Atlantic sturgeon
(*Acipenser oxyrinchus oxyrinchus*)
0 100 200 400 600 800
Kilometers
N
W
E
S

Figure 1

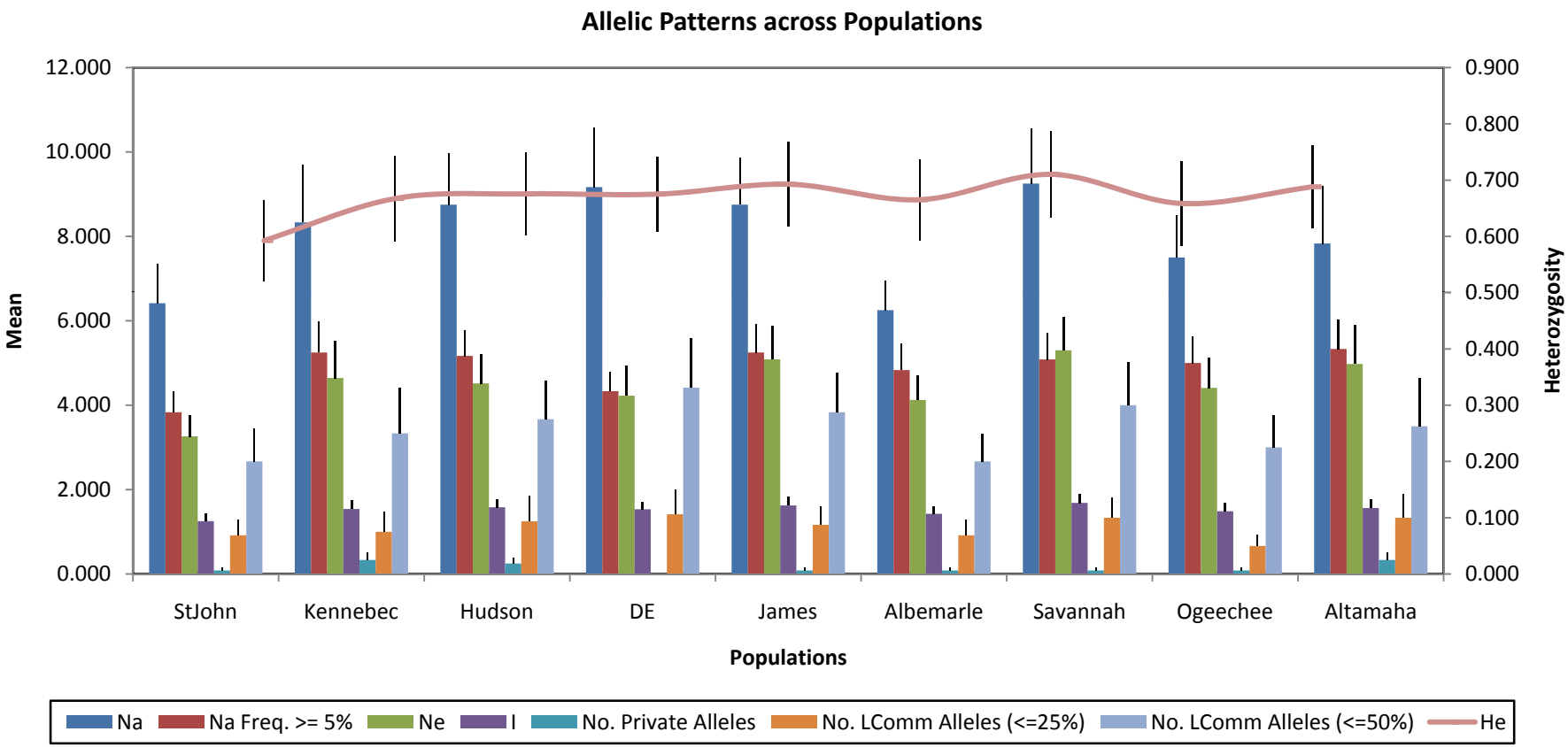
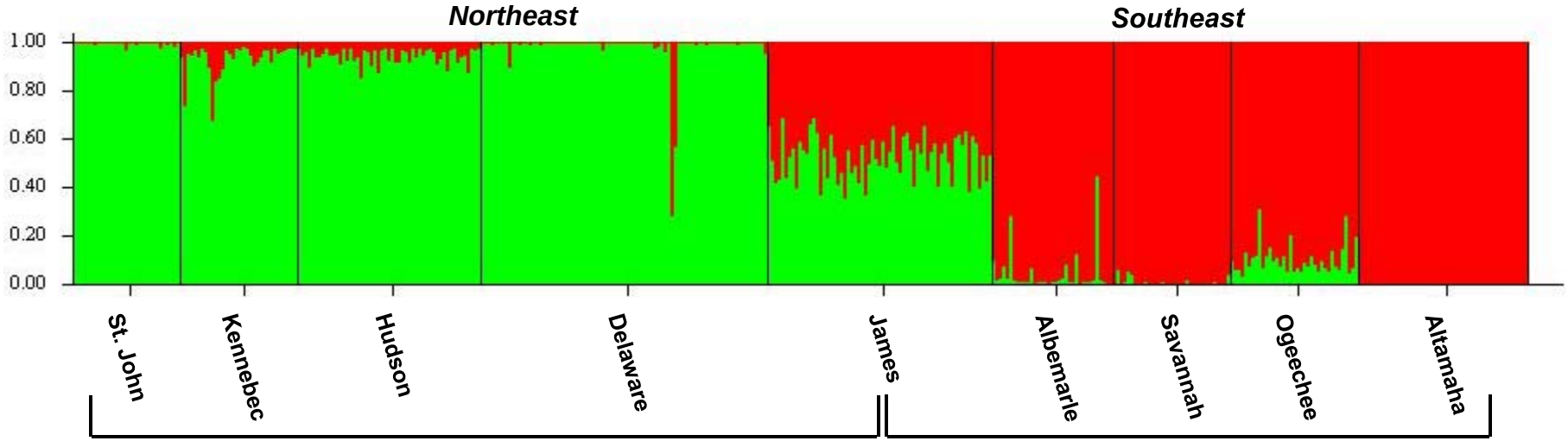

Allelic Patterns across Populations
12.000
10.000
8.000
6.000
4.000
2.000
0.000
Mean
0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
Heterozygosity
StJohn
Kennebec
Hudson
DE
James
Albemarle
Savannah
Ogeechee
Altamaha
Populations
Na
Na Freq. >= 5%
Ne
I
No. Private Alleles
No. LComm Alleles (<=25%)
No. LComm Alleles (<=50%)
He

Figure 2


Northeast
Southeast
1.00
0.80
0.60
0.40
0.20
0.00
St. John
Kennebec
Hudson
Delaware
James
Albemarle
Savannah
Ogeechee
Altamaha

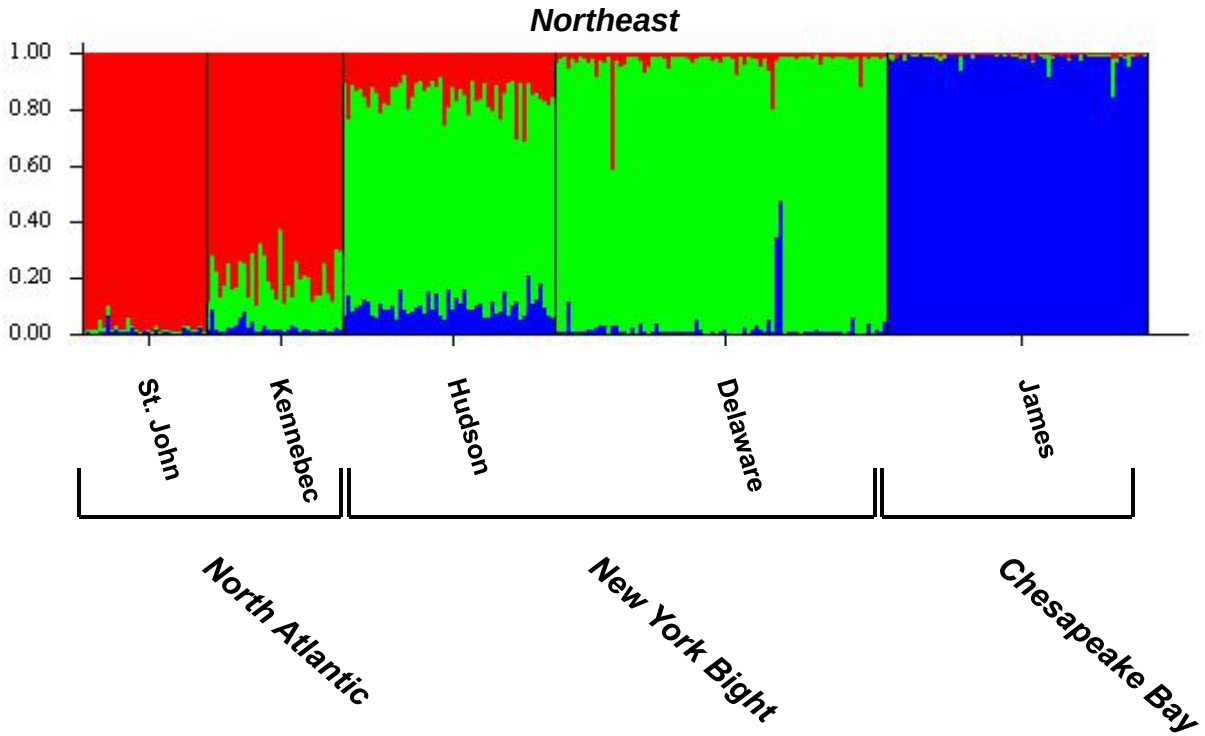

Northeast
1.00
0.80
0.60
0.40
0.20
0.00
St. John
Kennebec
Hudson
Delaware
James
North Atlantic
New York Bight
Chesapeake Bay

Figure 3

Multidimensional Scaling Plot of the Proportional of Shared Alleles Distances

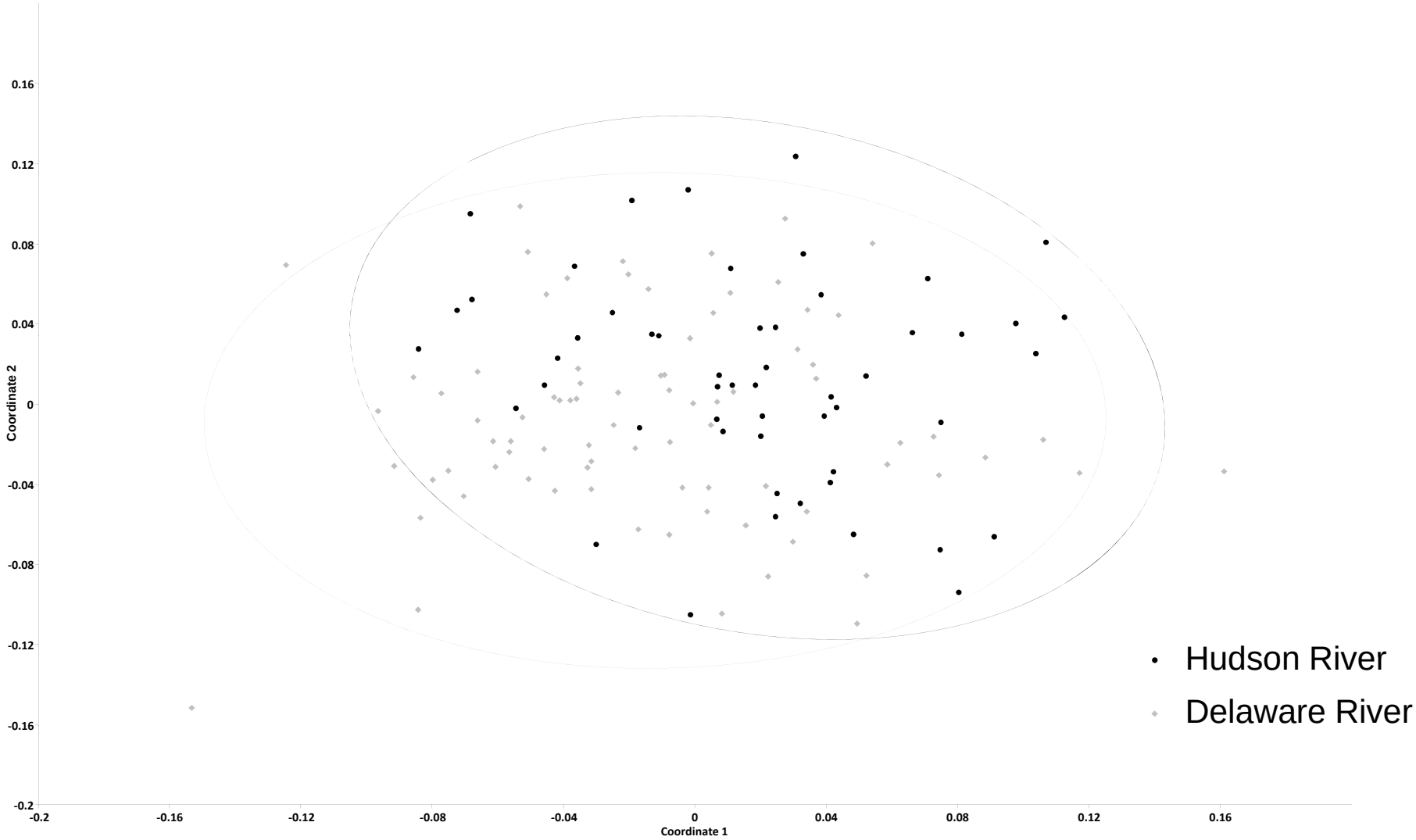


Figure 4

Multidimensional Scaling Plot of the Proportional of Shared Alleles Distances

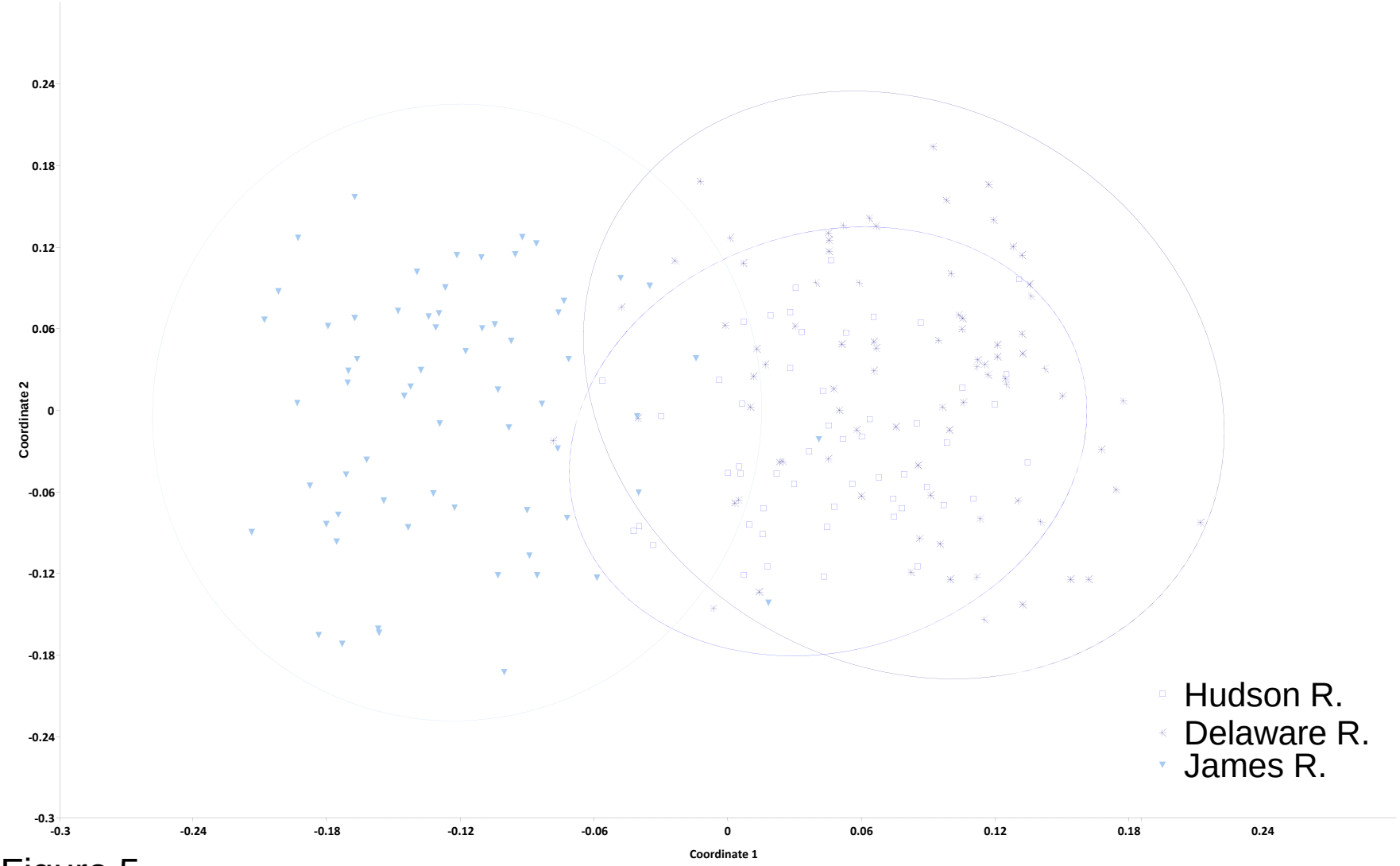
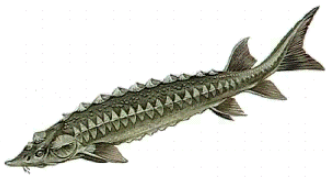
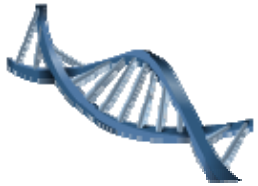
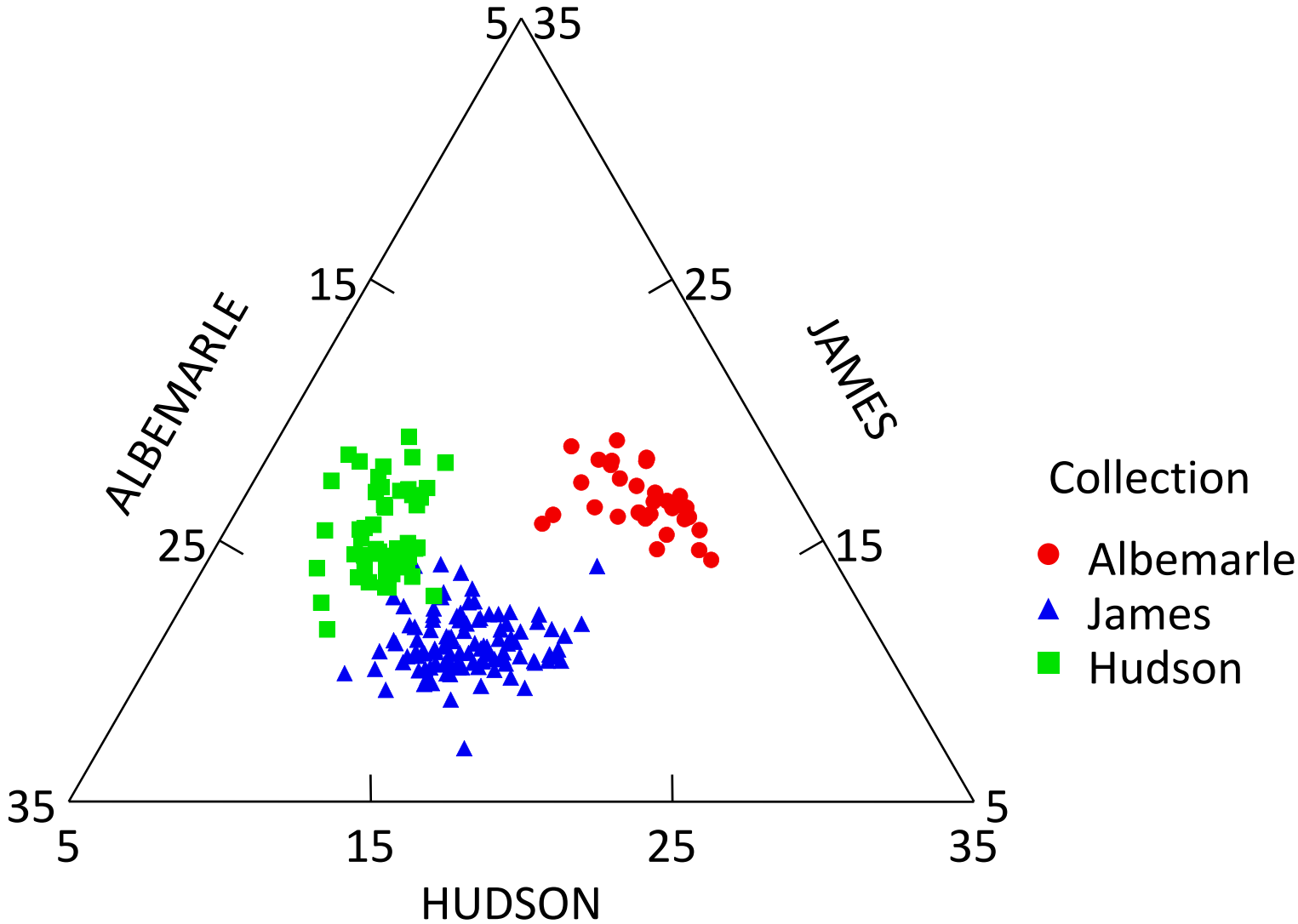


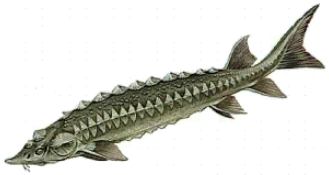
Figure 5



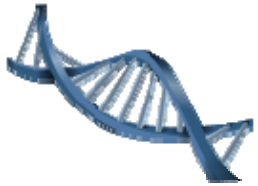
James River Comparison

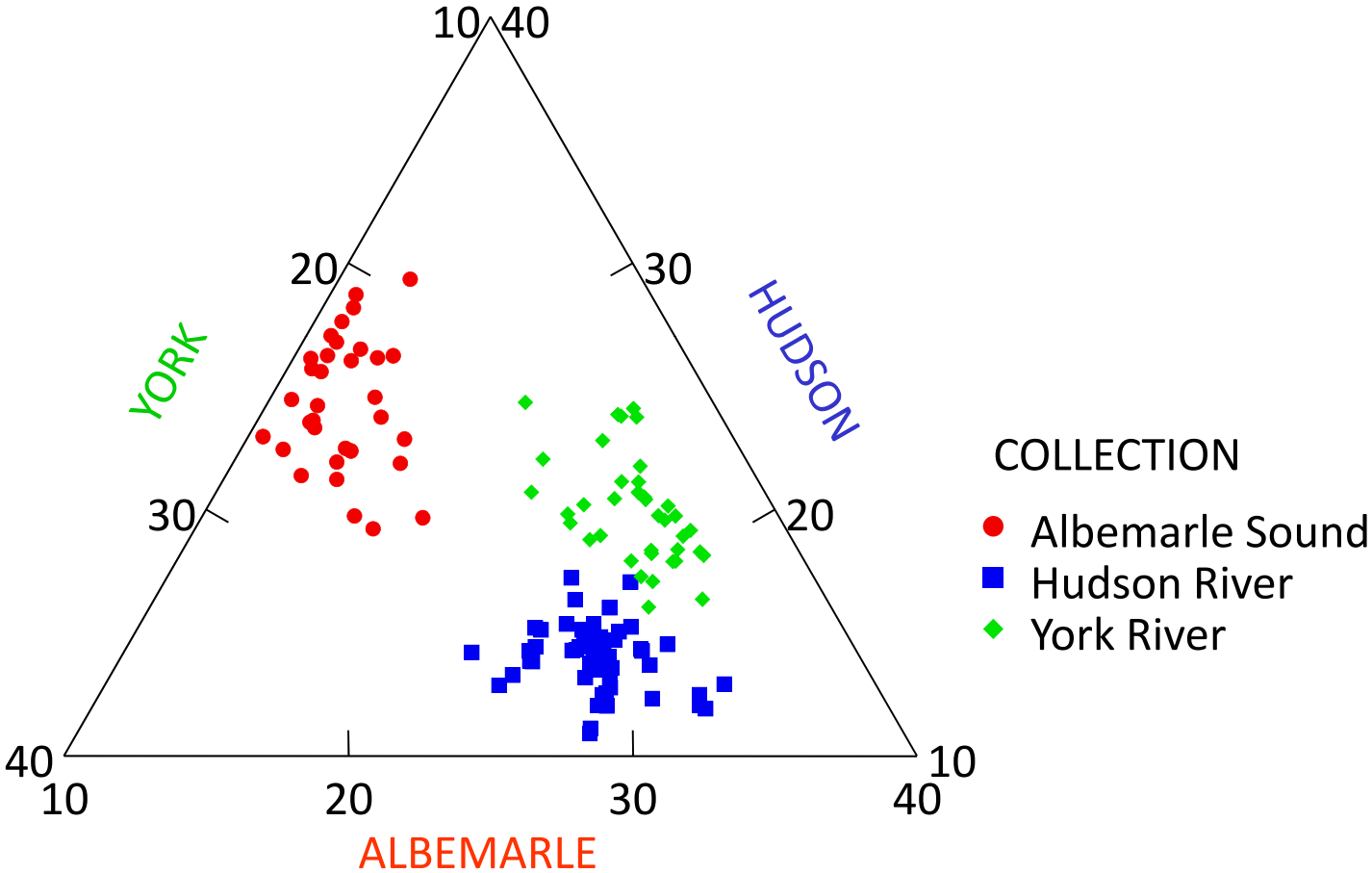




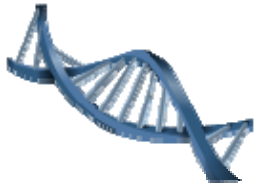


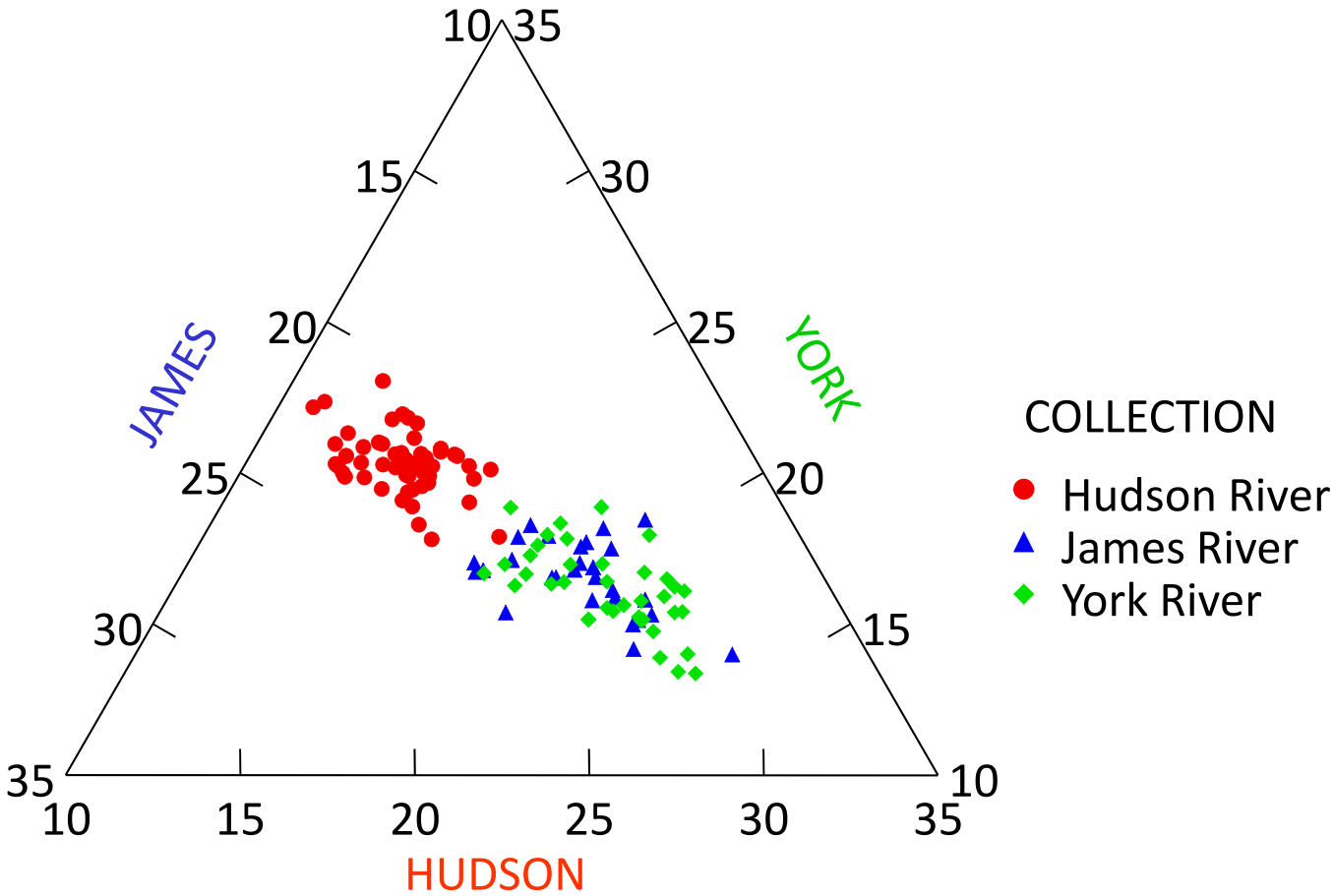
York River Comparison



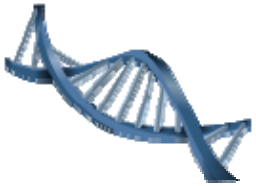


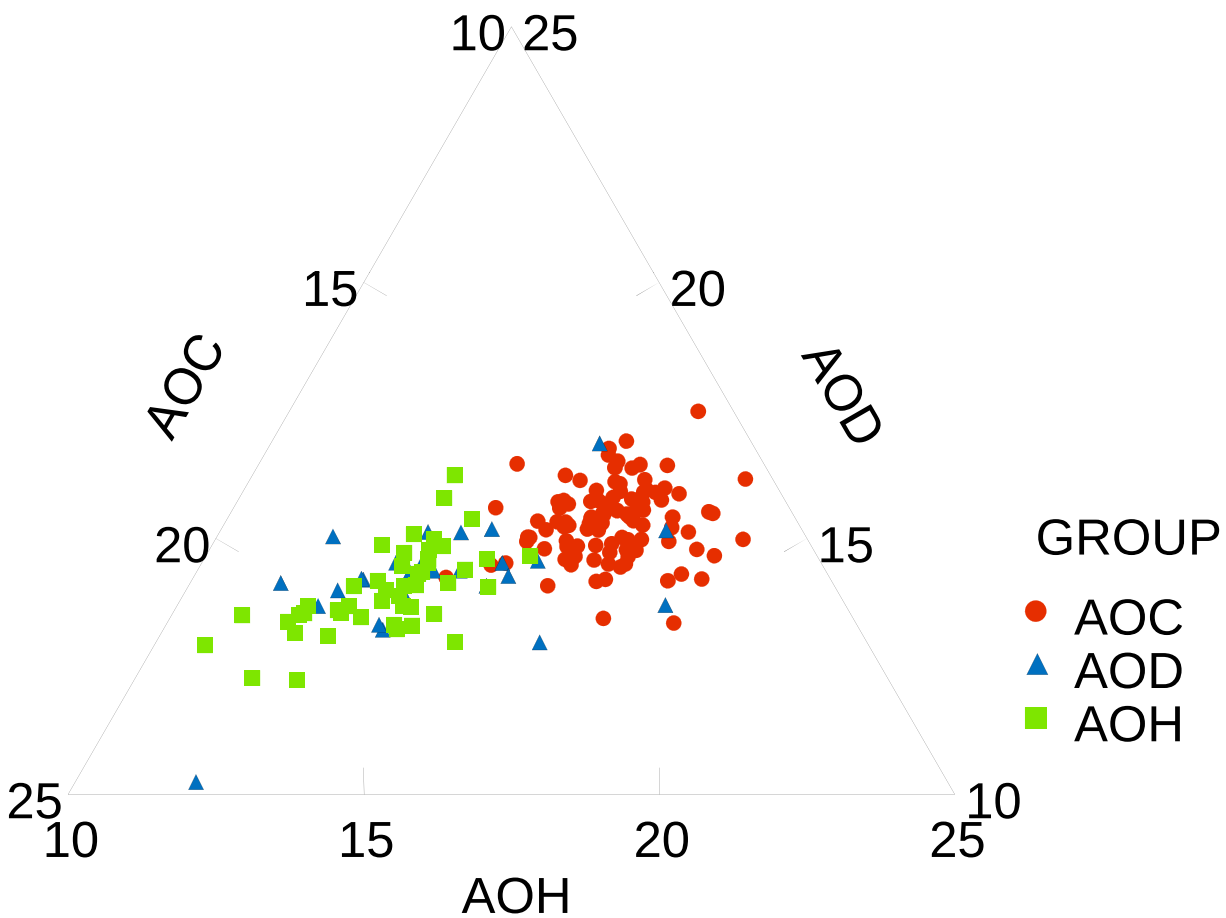
Hudson/James/York River Comparison

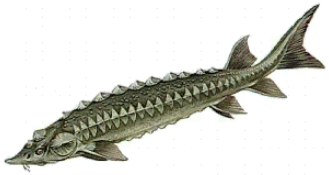




Hudson/Delaware/James River Comparison

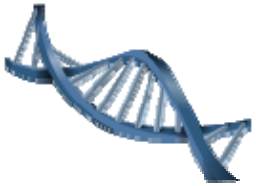


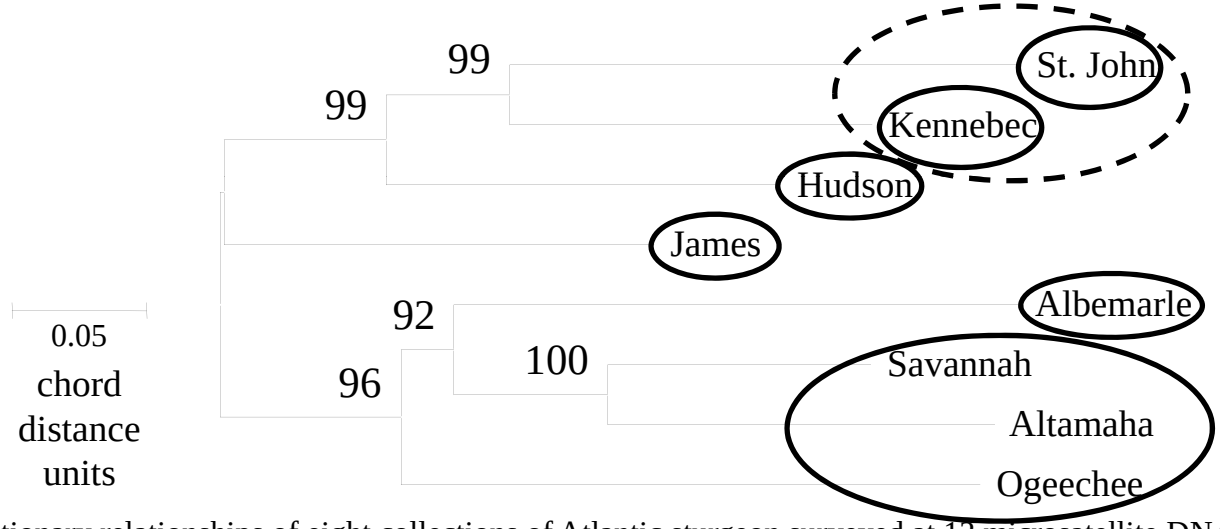




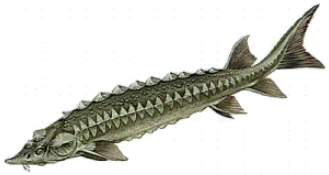
Chord Distance

N-J Tree

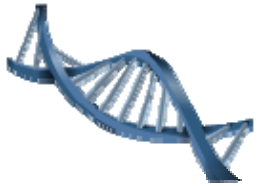




The evolutionary relationships of eight collections of Atlantic sturgeon surveyed at 12 microsatellite DNA markers inferred using the Neighbor-Joining method. The optimal tree, with bootstrap support after 5,000 permutations, is shown. Tree is drawn to scale with the branch lengths in chord distance units.



Assignment Testing



Collection	St. John	Kennebec	Hudson	James	Albemarle	Savannah	Ogeechee	Altamaha	Assign to Collection	Collection
St. John	29	2							93.5%	St. John
Kennebec	3	28	3						82.4%	Kennebec
Hudson		3	49	1					92.5%	Hudson
James			1	105					99.1%	James
Albemarle			1		33	1			94.3%	Albemarle
Savannah				1	2	22		9	64.7%	Savannah
Ogeechee						2	35		94.6%	Ogeechee
Altamaha						7	1	41	83.7%	Altamaha
Assign to DPS	93.5%	82.4%	92.5%	99.1%	94.3%	97.5%			90.2% / 95.3%	
DPS	Canadian	GOM	NY Bight	Chesapeake	Carolina	South Atlantic				

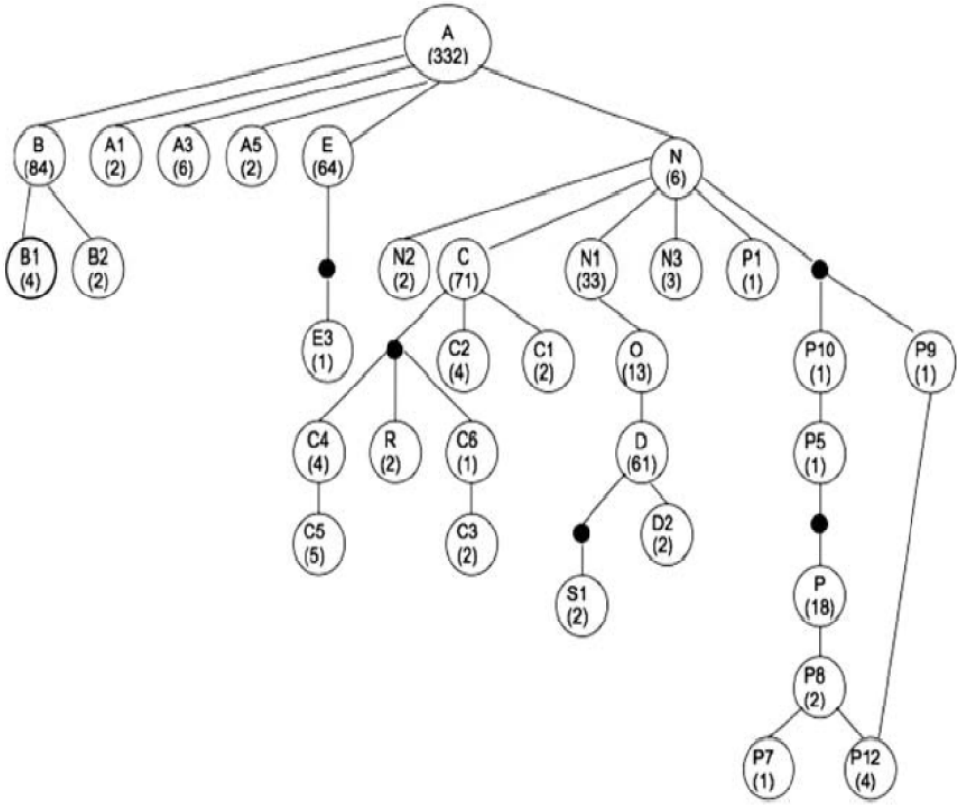
Overall Assignment Success: Collection- 90.2% DPS- 95.3%

Incorrect assignments are distributed horizontally and are highlighted in yellow.

Grunwald et al. 2007

Haplotype Network

Fig. 2 Network of mtDNA control region sequence haplotypes of Atlantic sturgeon developed using statistical parsimony as implemented in TCS vs. 1.3.1 (Clement et al. 2000). Numbers in parentheses indicate the number of individuals with that haplotype. Black dots represent unobserved haplotypes and haplotypes connected by a single line differ by one nucleotide



UPGMA Tree of Genetic Distances

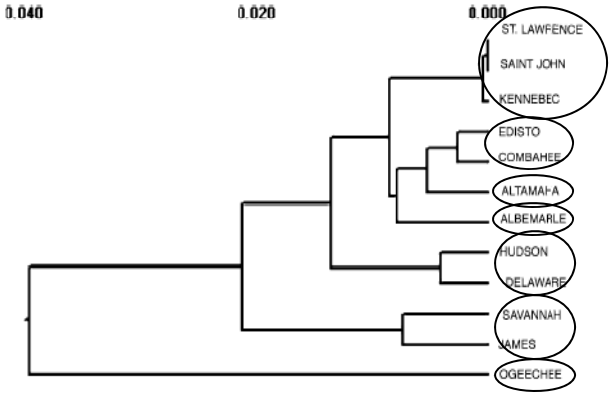


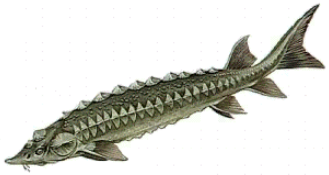
Fig. 3 An UPGMA tree of the population genetic distances for the mtDNA control region sequence data from Atlantic sturgeon of 12 Atlantic coast rivers and estuaries

UPGMA tree

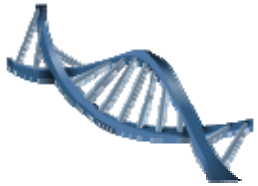
(1) St. Lawrence, Saint John, and Kennebec, (2) Hudson and Delaware, (3) James and Savannah, (4) Albemarle, (5) Edisto and Combahee, (6) Ogeechee, and (7) Altamaha rivers.

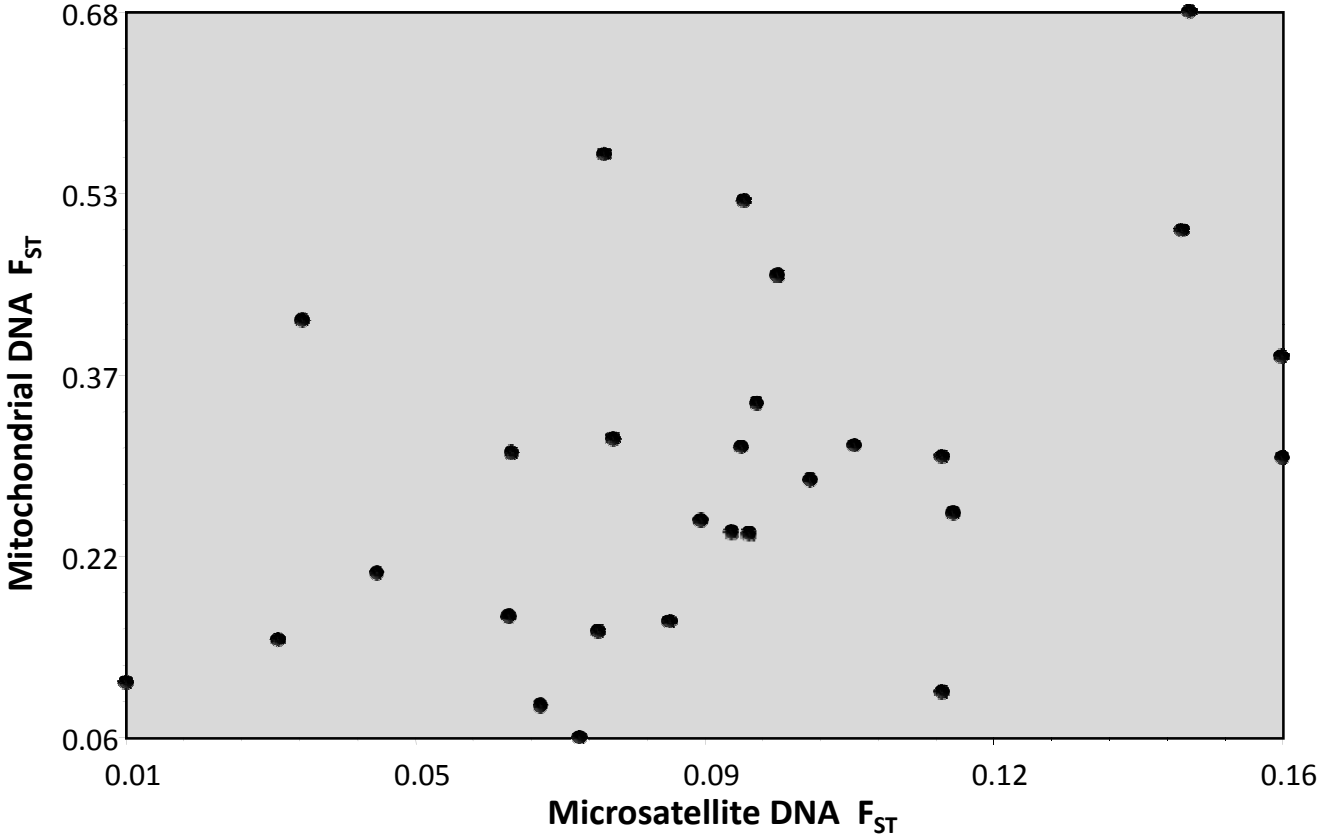
AMOVA

(1) St. Lawrence, St. John, and Kennebec, (2) Hudson,
(3) Delaware, (4) James, (5) Albemarle,
(6) Combahee and Edisto, (7) Savannah, (8) Ogeechee, and
(9) Altamaha



Mantel Test
Comparing Estimates of Pair-wise Population Differentiation (F_{ST}) between mtDNA and nDNA



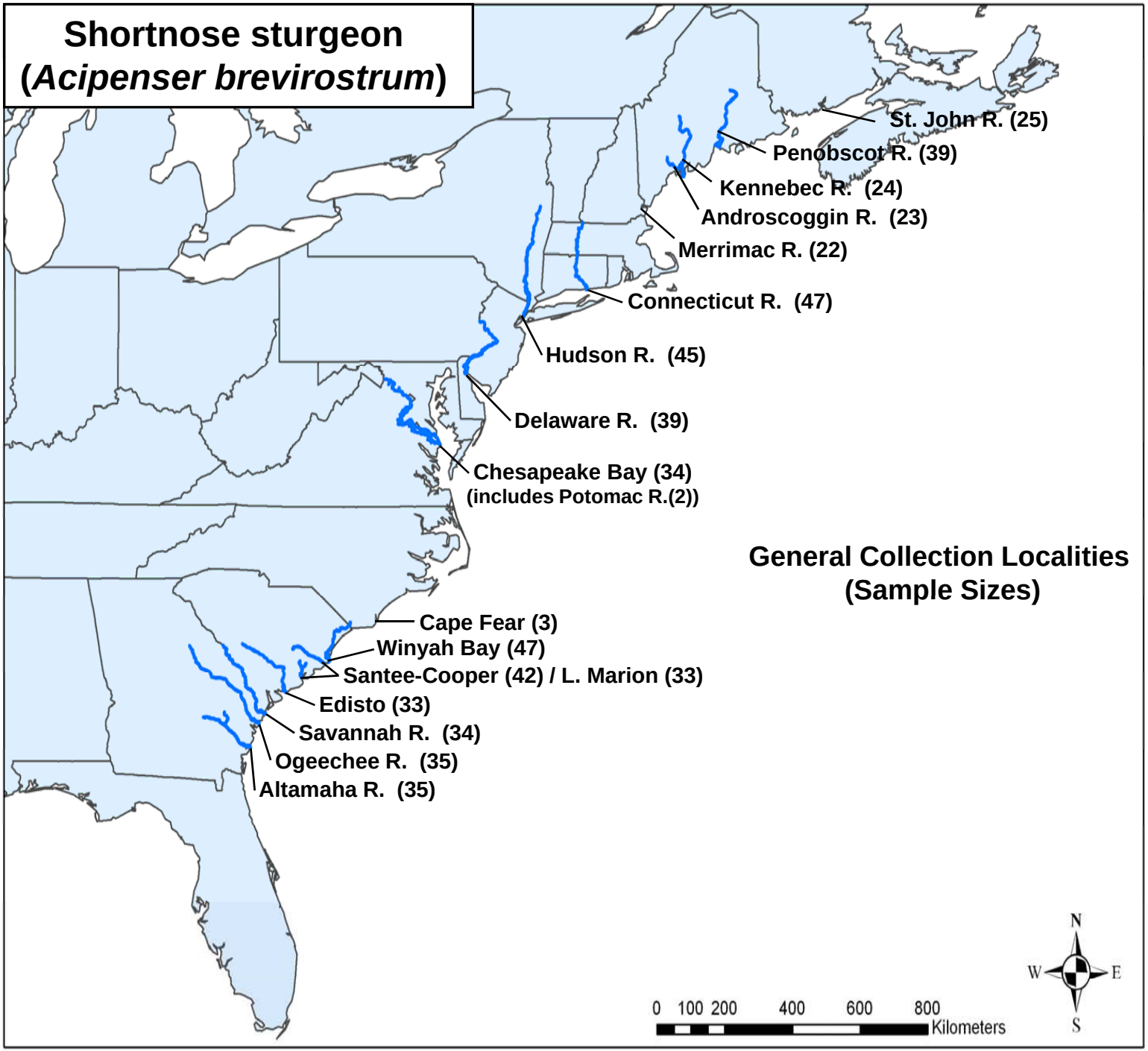


Scatterplot depicting the Mantel analysis comparing the mtDNA F_{ST} values for eight Atlantic coast collections of Atlantic sturgeon and the pair-wise microsatellite DNA F_{ST} values for the same collections surveyed at 12 microsatellite DNA loci (Mantel test: $r = 0.45$; $t = 2.25$; $P = 0.021$).

Shortnose Sturgeon

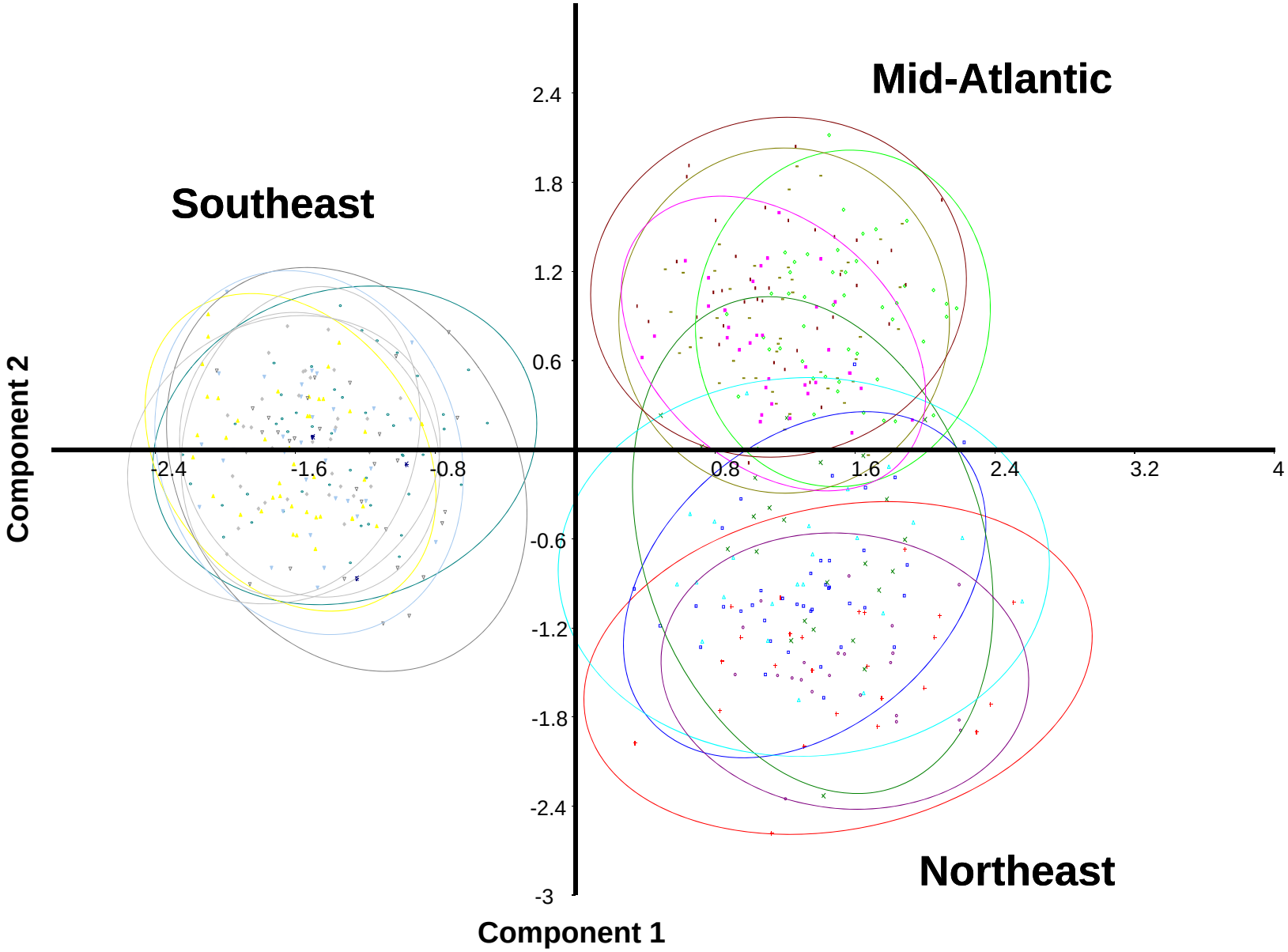
Overview

- *A. brevirostrum* listed under ESPA (predecessor to ESA) in 1967
- 1998 recovery plan recommended that all 19 estuaries should be managed as DPSs (no genetics information available at the time); status currently under review;
- 96 polysomic microsatellite DNA markers have been developed *for A. brevirostrum*
  - Markers tested for inheritance among 7 families
- No markers have been found that appear to be disomic
- 11 markers surveyed in 561 fish sampled from 17 rivers/embayments
- A series of analyses have been performed in which the allelic patterns observed at these these loci were treated as dominant (i.e., presence/absence) markers
- A dataset of 561 animals surveyed for 220 alleles was generated
- Unique alleles observed in many collections
- Comparison between distance measures – Jaccard vs Φ_{PT}
  - Mantel test indicates the two distance are nearly identical ($r=0.98$; $P<0.0001$)
- Other analyses: PCA, ANOSIM, N-J cluster, STRUCTURE, AMOVA, assignment,

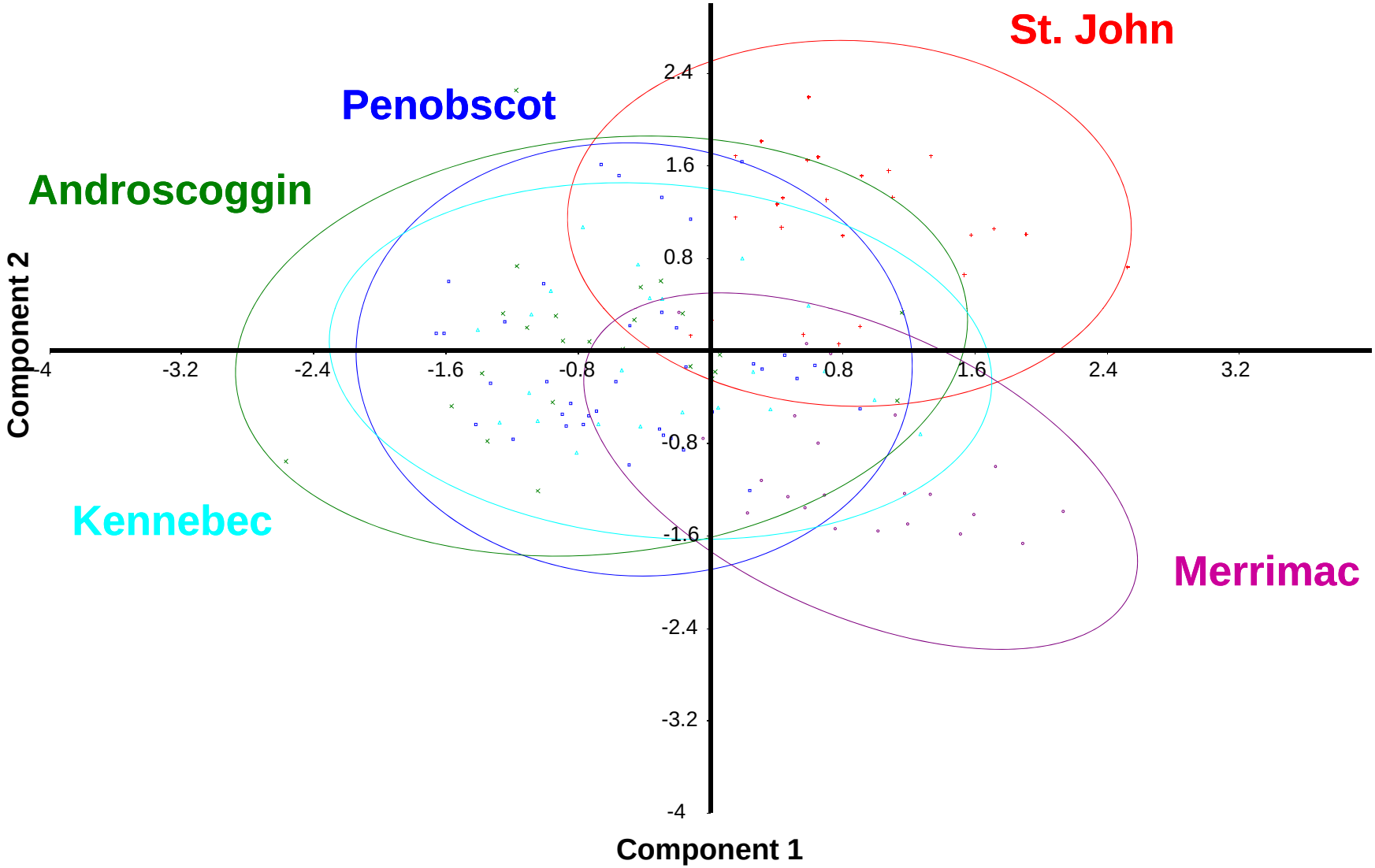

Shortnose sturgeon
(*Acipenser brevirostrum*)
St. John R. (25)
Penobscot R. (39)
Kennebec R. (24)
Androscoggin R. (23)
Merrimac R. (22)
Connecticut R. (47)
Hudson R. (45)
Delaware R. (39)
Chesapeake Bay (34)
(includes Potomac R.(2))
General Collection Localities
(Sample Sizes)
Cape Fear (3)
Winyah Bay (47)
Santee-Cooper (42) / L. Marion (33)
Edisto (33)
Savannah R. (34)
Ogeechee R. (35)
Altamaha R. (35)
0 100 200 400 600 800
Kilometers
N
W
E
S

Range-wide PCA

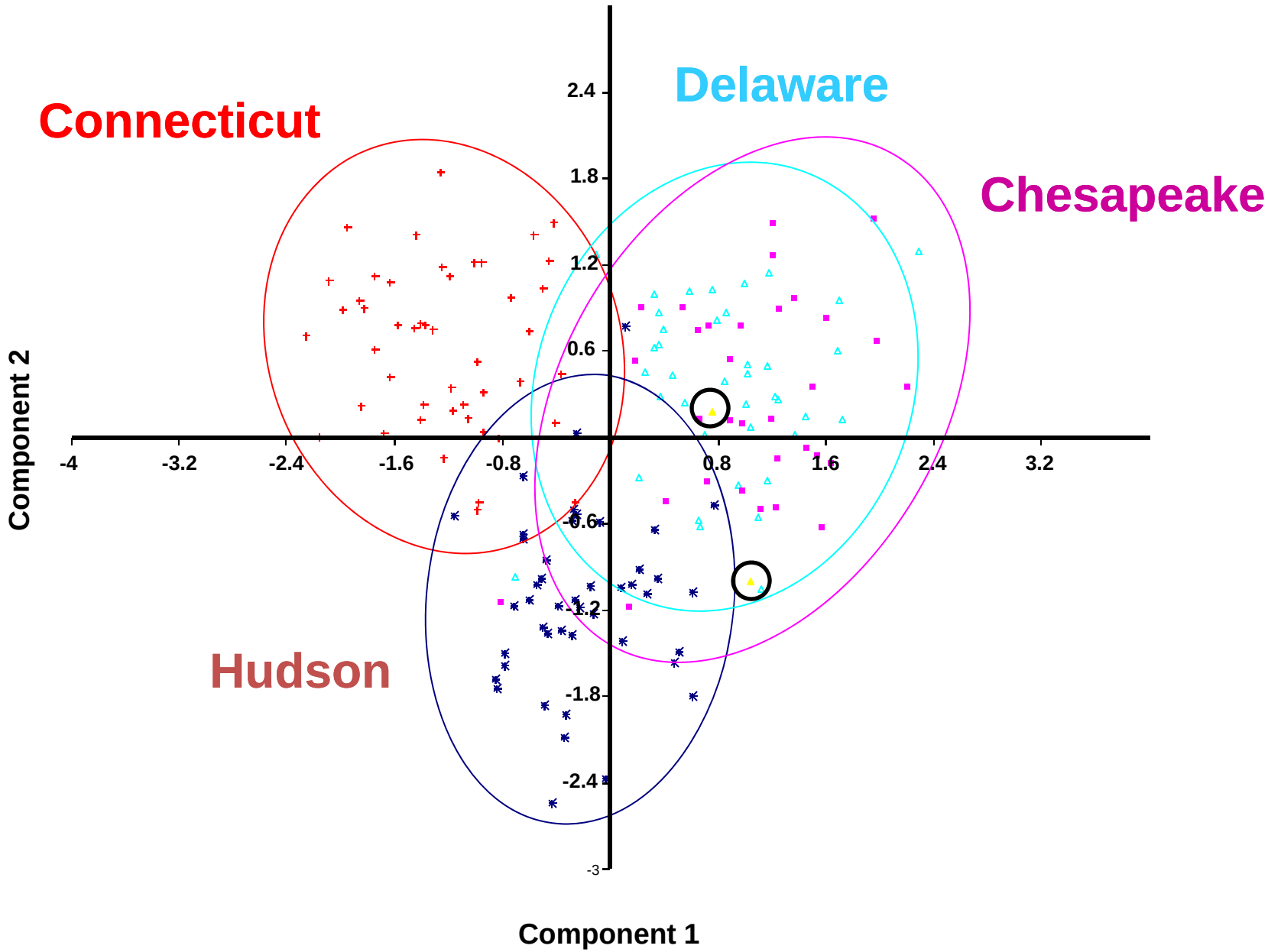
with 95% ellipses



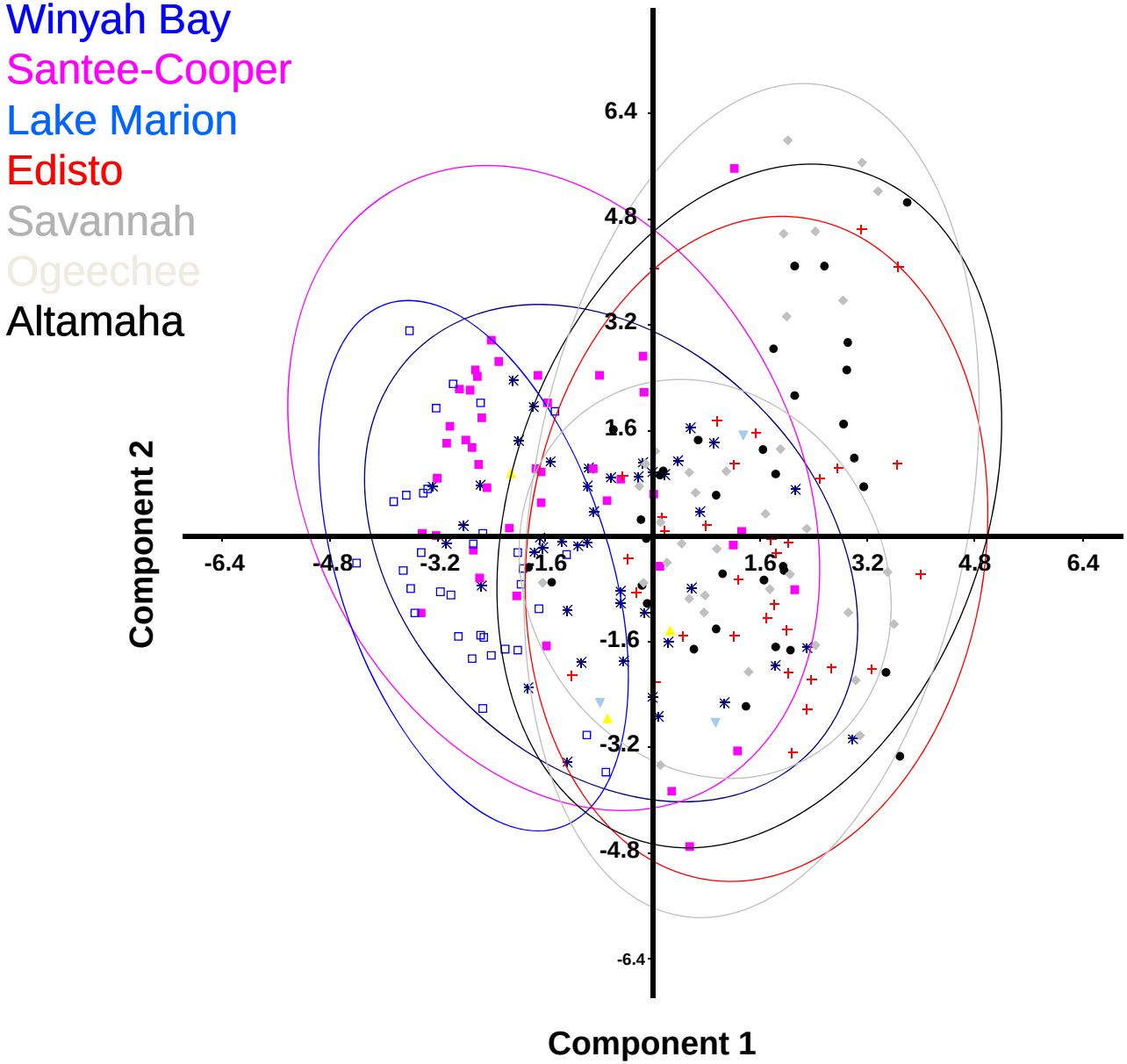
Northeast PCA



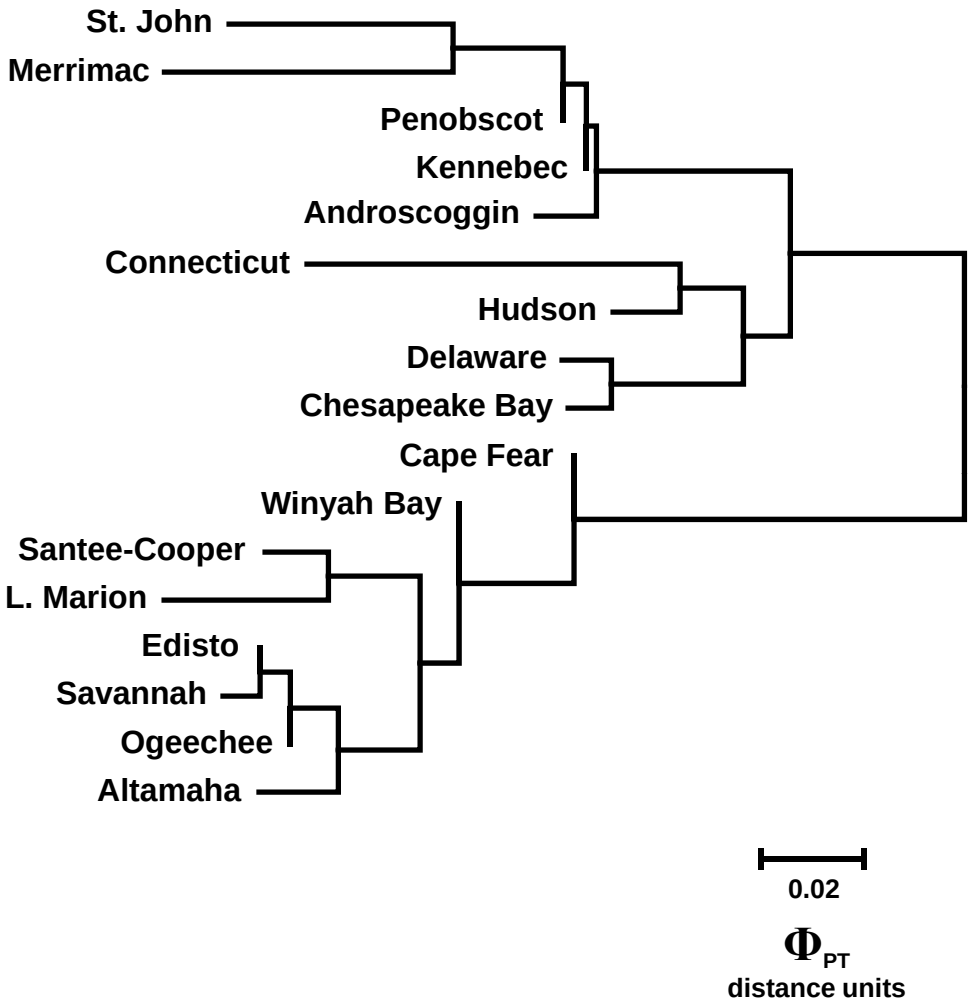
Mid-Atlantic PCA



Southeast PCA



N-J Tree

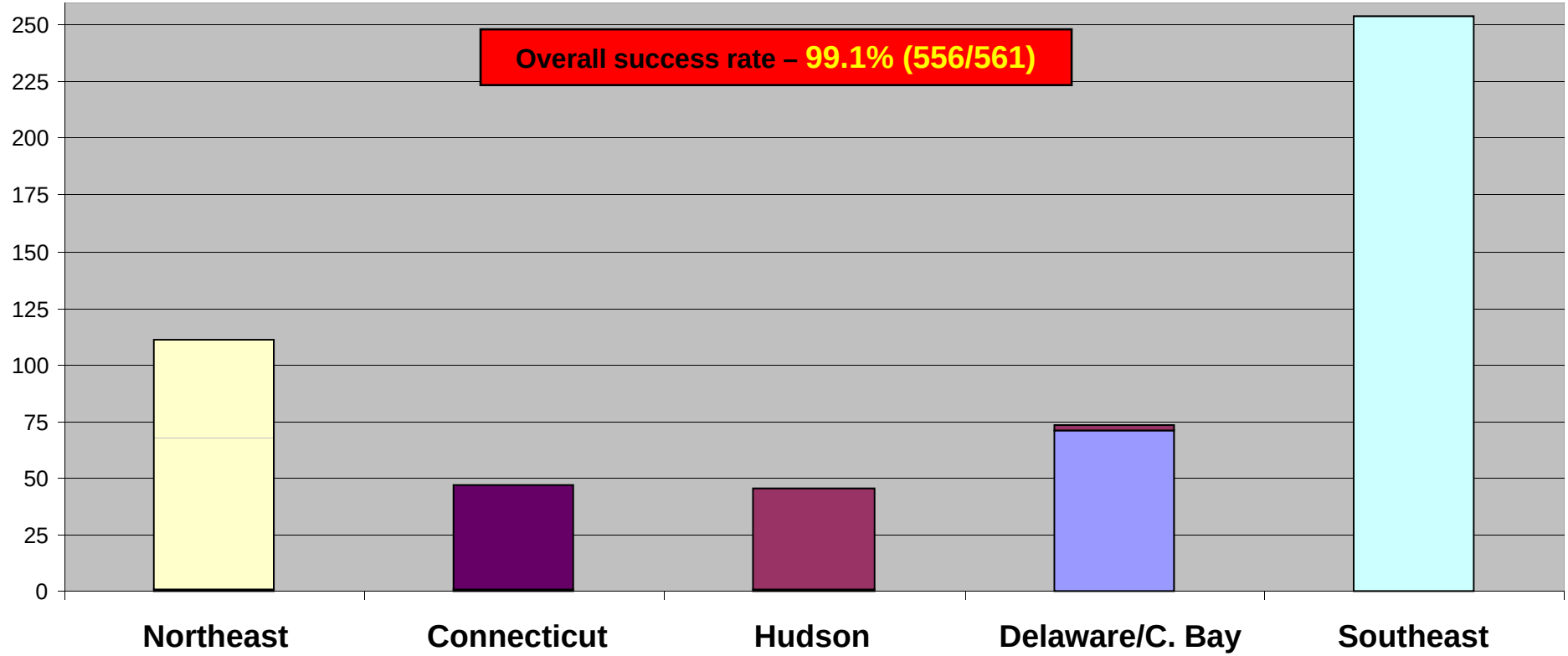


Evolutionary relationships of 17 taxa

The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree with the sum of branch length = 0.41323750 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Phylogenetic analyses were conducted in MEGA4 [2].

1. Saitou N & Nei M (**1987**) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406-425.
2. Tamura K, Dudley J, Nei M & Kumar S (**2007**) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **10.1093/molbev/msm092**.

Allocation to DPS - 5 Segment Model



allocated to	Northeast	Connecticut	Hudson	Delaware/C. Bay	Southeast
Northeast	132	0	0	0	0
Connecticut	0	46	0	0	0
Hudson	0	1	44	2	0
Delaware/C. Bay	1	0	1	71	0
Southeast	0	0	0	0	254
Assignment %	99.30	97.90	97.80	97.30	100.00

The mtDNA Perspective
(Wirgin et al. 2005)

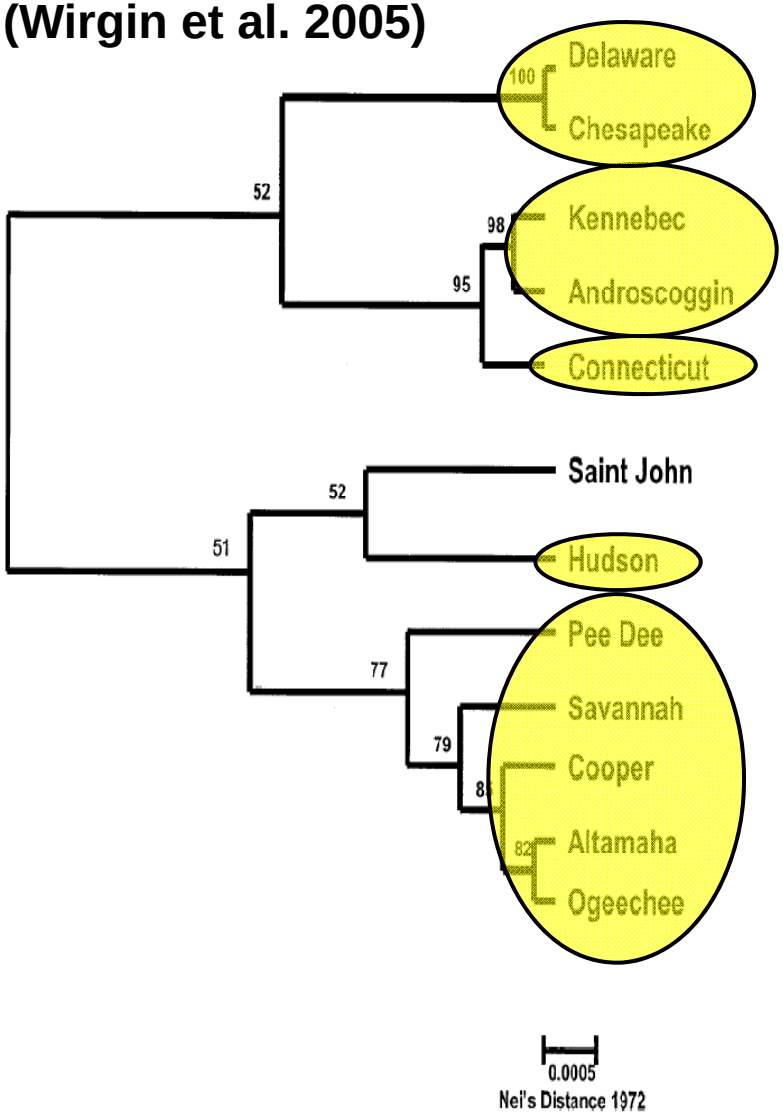
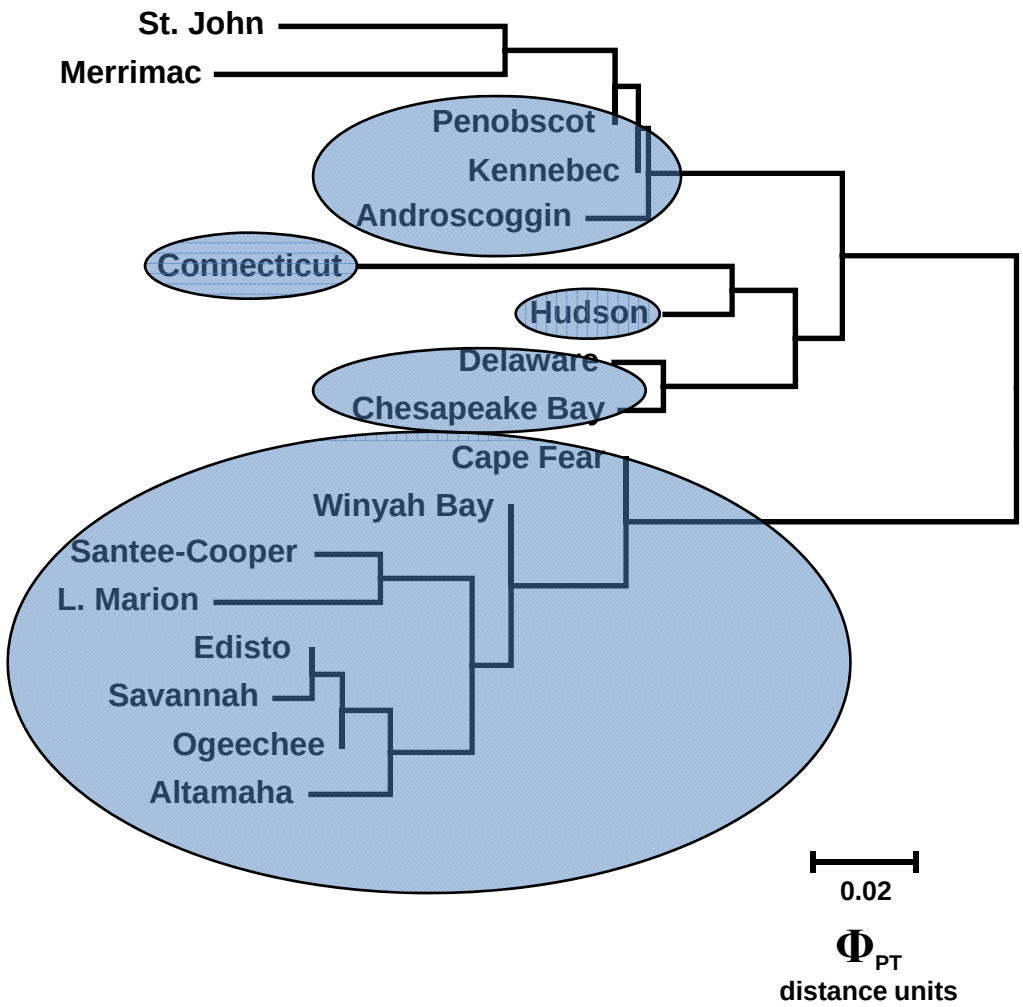


Fig. 2. An UPGMA tree of the population genetic distances for the mtDNA control region sequence data from shortnose sturgeon of 12 Atlantic Coast rivers and estuaries. Bootstrap values are given at each node.

The nDNA Perspective
5 DPS Model



Neighbor-Joining tree illustrating the structure in the Φ_{PT} distance matrix generated from 11 polysomic microsatellite DNA loci surveyed in *A. brevirostrum* from 17 Atlantic coast collections.

Summary

- Principle components analysis (PCA) of these data indicated the presence of 3 major (regional) groupings among the 17 surveyed river/bay systems. These regional groupings are:

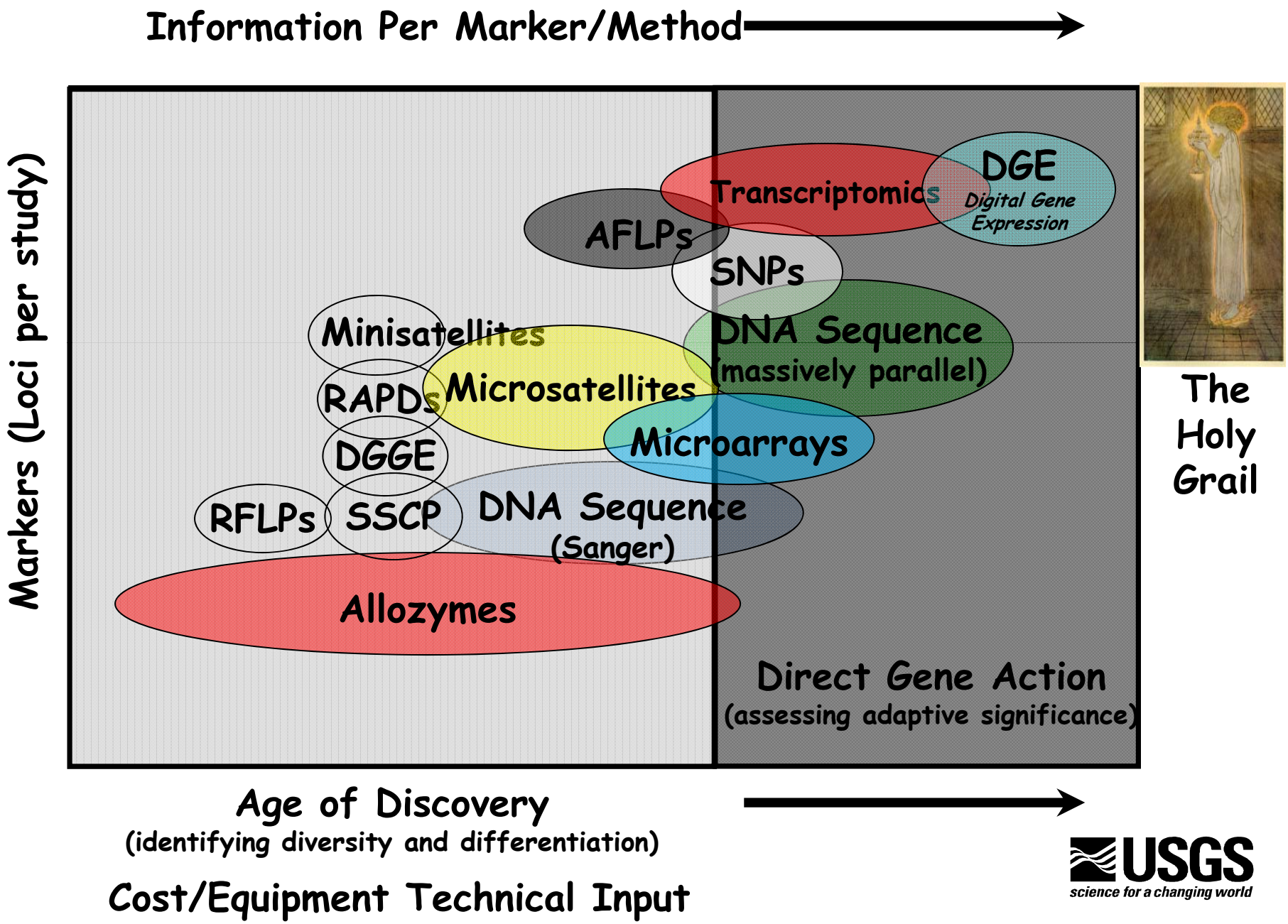
    1. ***Northeast*** which includes the rivers from the Gulf of Maine
    2. ***Mid-Atlantic*** which includes the CT, Hudson, and DE Rivers
    3. ***Southeast*** including the Cape Fear, Winyah Bay, Santee-Cooper system, Edisto, Savannah, Ogeechee, and Altamaha Rivers

- Regional PCA suggests the three ***Mid-Atlantic*** rivers may contain distinct populations of *A. brevirostrum;* the Southeastern PCA suggests much overlap; the northern- (Santee-Cooper) and southern-most (Altamaha) collections show the least overlap

- Hierarchical *STRUCTURE results were consistent with PCA analyses as* K (clusters) was estimated to be 5: three major clusters with the mid-Atlantic cluster consisting of three sub-clusters.

- Neighbor-Joining tree is consistent with PCA, STRUCTURE analysis

Next-Generation Conservation Biology – Comparative Population Genomics



USGS
science for a changing world

Newer Technologies

- Next Generation Sequencing
  - Massively parallel sequencing
    - Roche 454
    - Applied Biosystems SOLID system
    - Solexa's Illumina Genome Analyzer

Next-generation DNA sequencing

Jay Shendure[1] & Hanlee Ji[2]

DNA sequence represents a single format onto which a broad range of biological phenomena can be projected for high-throughput data collection. Over the past three years, massively parallel DNA sequencing platforms have become widely available, reducing the cost of DNA sequencing by over two orders of magnitude, and democratizing the field by putting the sequencing capacity of a major genome center in the hands of individual investigators. These new technologies are rapidly evolving, and near-term challenges include the development of robust protocols for generating sequencing libraries, building effective new approaches to data-analysis, and often a rethinking of experimental design. Next-generation DNA sequencing has the potential to dramatically accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes, transcriptomes and interactomes to become inexpensive, routine and widespread, rather than requiring significant production-scale efforts.

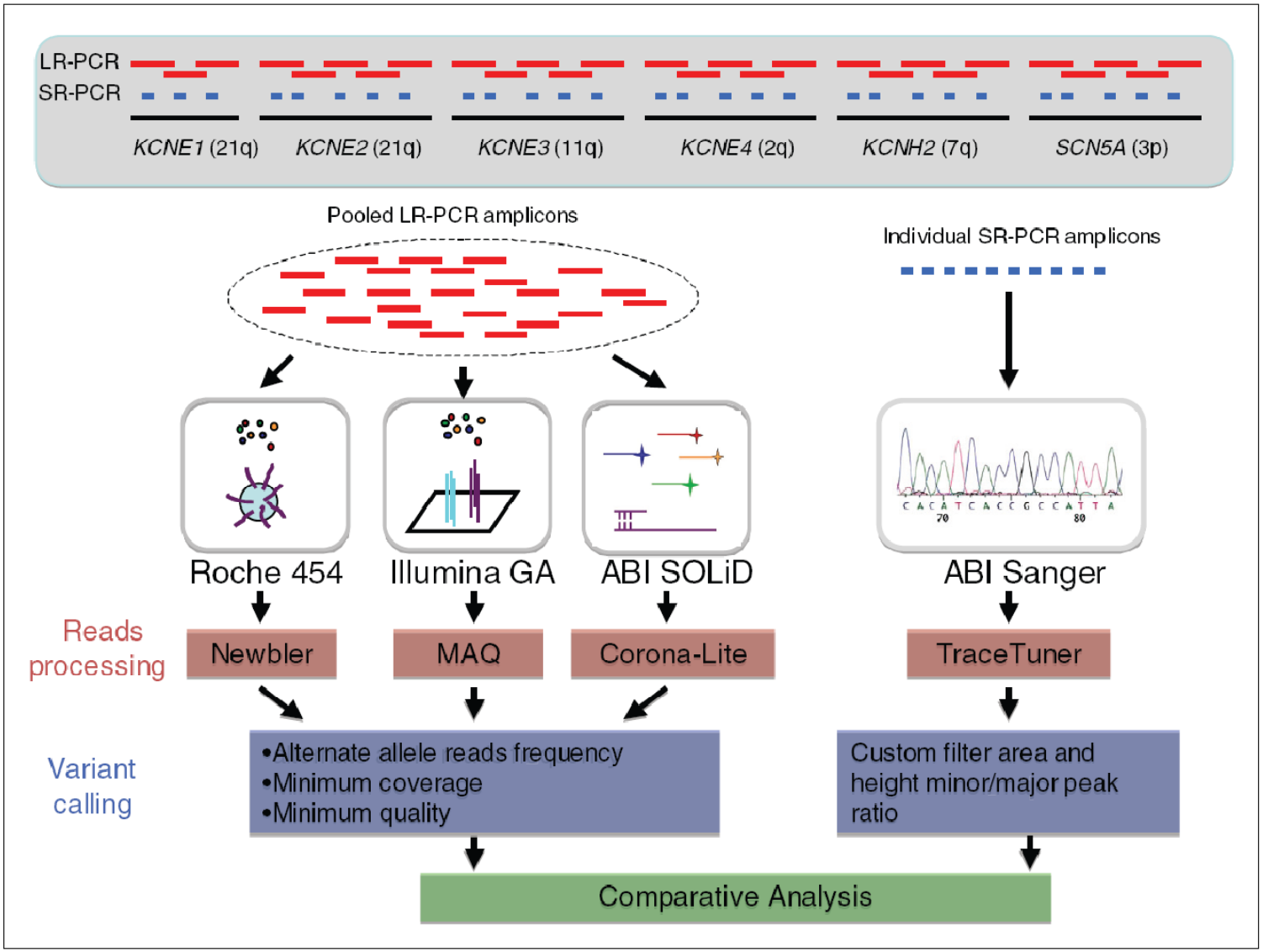


Figure 1

Overview of experimental design. Six genomic intervals, each encoding genes for K^+/Na^+ voltage-gated channel proteins, were amplified using DNA from four individuals and LR-PCR reactions to generate 260 kb of target sequence per sample. Amplicons from each individual were pooled in equimolar amounts and then sequenced using the three NGS platforms. The 260 kb examined in this study is representative of human sequences containing 38% repeats and 4% coding sequence compared with 47% and 1%, respectively, genome-wide. For each sample 88 kb was amplified using short range PCR (SR-PCR) reactions targeting the exons and evolutionarily conserved intronic regions. Each SR-PCR amplicon was individually sequenced in the forward and reverse directions using the ABI-3730xL platform (Additional data file 2). Data generated from the NGS platforms were analyzed to identify bases variants from the reference sequence (build 36) and the quality of the variant calls was assessed using platform specific methodologies. A comparative analysis of the sequence data from the NGS platforms and ABI Sanger was then performed to determine accuracy, and false positive and false negative rates.

Genomics of the Acipenseridae

Transcriptome Assembly

All RNA molecules, including mRNA, rRNA, tRNA, and non-coding RNA

The term can be applied to the total set of transcripts in a given organism, or to the specific subset of transcripts present in a particular cell type.

Marker Development

Microsatellite DNA
Single Nucleotide Polymorphisms (SNPs)

Mitogenomics

Study of the complete mitochondrial genome; More accurately assess phylogenetics, phylogeography and estimate divergence times

Genomics of the Acipenseridae

Transcriptome Assembly >2,000,000 reads

Acipenser oxyrinchus
Acipenser brevirsotrum
Scaphirhynchus albus
Scaphirhynchus platorynchus

Marker Development

Microsatellite DNA > 66,000 potential markers
SNP > 70,000 potential markers

Genomics of the Acipenseridae

Transcriptome Assembly >2,000,000 reads

Acipenser oxyrinchus
Acipenser brevirsotrum
Scaphirhynchus albus
Scaphirhynchus platorynchus

Marker Development

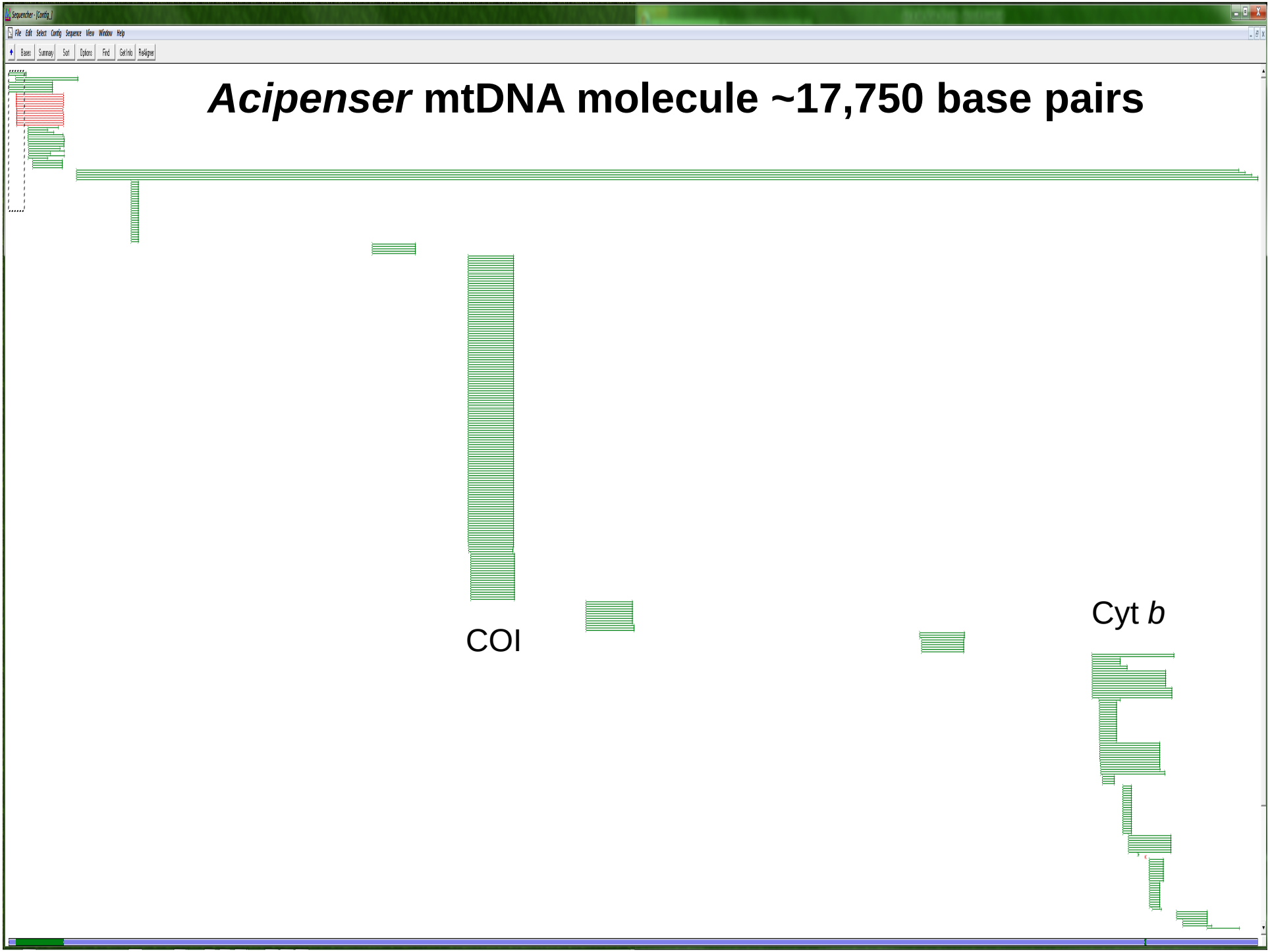
Microsatellite DNA > 66,000 potential markers
SNPs > 70,000 potential markers

Mitogenomics

Five complete *Acipenser* mtDNA genomes are in GenBank (*A. gueldenstaedtii, A. dabryanus, A. sinensis, A. stellatus*, and *A. transmontanus*)

Compiled a contig of 375 sequences from GenBank and new sequences resulting from transcriptome assembly

Acipenser mtDNA molecule ~17,750 base pairs

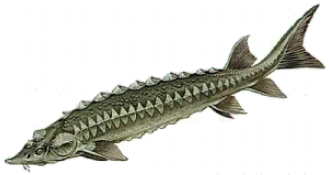


Mitogenomics

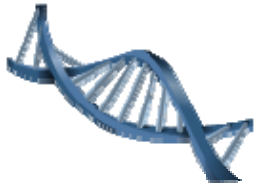
Long-range PCR of two regions (COI and Cyt b)

Long-read massively parallel sequencing (e.g., 454 Roche) for the sequence, assemble, and phylogenetic comparison

Coding genes of the mtDNA may evolve under selection owing to metabolic requirements.



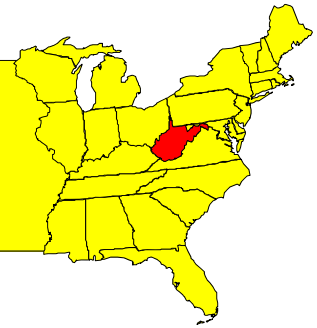
Acknowledgements



Funding provided by

NOAA Fisheries and USGS Leetown Science Center

Thank you to Ike Wirgin, Doug Peterson, Kim Damon-Randall, Wes Patrick, Stepania Bolden, Wilson Laney, Joseph Hightower, Steve Minkkinen, Jerre Mohler, Ken Sulak, Kevin Friedland, Steve Mierzykowski, Mark Collins, Andy Kahnle, for providing samples and/or discussions related to the collection and

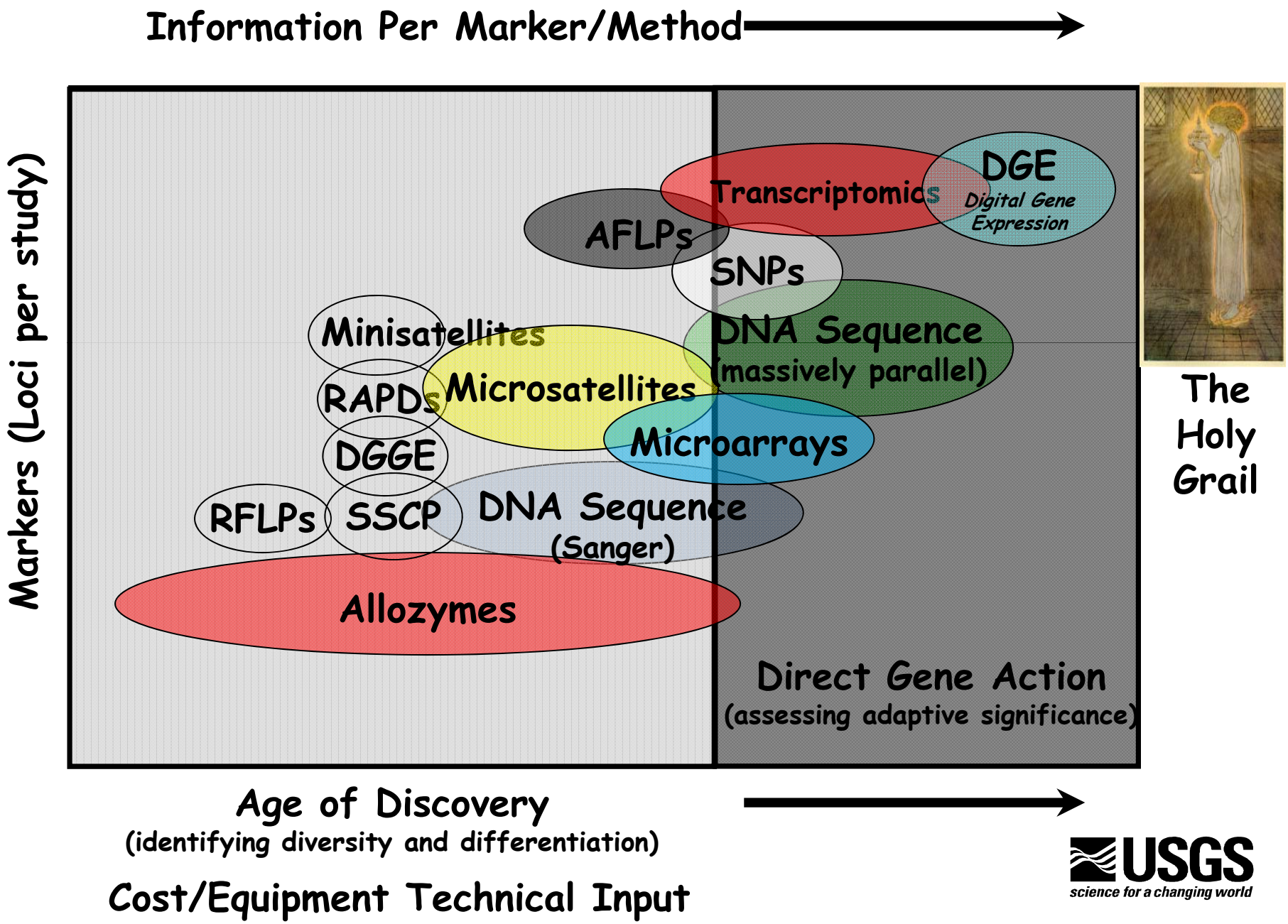


Functional Genomics

Goal: Use powerful tools (e.g., RNA-seq) to investigate the evolutionary processes of adaptive divergence among natural populations.

Research Question - Does the genetic divergence observed among wild population observed with neutral markers show adaptive significance (i.e., or is differentiation due to divergent natural selection)?

Next-Generation Conservation Biology – Comparative Population Genomics



USGS
science for a changing world

Next Generation Sequencing

Massively parallel sequencing

Table 1 Second-generation DNA sequencing technologies

	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length	References
454	Emulsion PCR	Polymerase (pyrosequencing)	~$60	$500,000	Yes	Indel	250 bp	14,20
Solexa	Bridge PCR	Polymerase (reversible terminators)	~$2	$430,000	Yes	Subst.	36 bp	17,22
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~$2	$591,000	Yes	Subst.	35 bp	13,26
Polonator	Emulsion PCR	Ligase (nonamers)	~$1	$155,000	Yes	Subst.	13 bp	13,20
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~$1	$1,350,000	Yes	Del	30 bp	18,30

The pace with which the field is moving makes it likely that estimates for costs and read-lengths will be quickly outdated. Vendors including Roche Applied Science, Illumina, and Applied Biosystems have major upgrade releases currently in progress. Estimated costs-per-megabase are approximate and inclusive only of reagents. Read-lengths are for single tags. Subst., substitutions; indel, insertions or deletions; del, deletions.

Table 2 Applications of next-generation sequencing

Category	Examples of applications	Refs
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes	44
Reduced representation sequencing	Large-scale polymorphism discovery	45
Targeted genomic resequencing	Targeted polymorphism and mutation discovery	46–52
Paired end sequencing	Discovery of inherited and acquired structural variation	53,54
Metagenomic sequencing	Discovery of infectious and commensal flora	55
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations	56–63
Small RNA sequencing	microRNA profiling	64
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA	60,65,66
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions	67–70
Nuclease fragmentation and sequencing	Nucleosome positioning	69
Molecular barcoding	Multiplex sequencing of samples from multiple individuals	61,71

De novo Transcriptome Development

NEWS AND VIEWS

PERSPECTIVE

Sequencing goes 454 and takes large-scale genomics into the wild

HANS ELLEGREN
Department of Evolutionary Biology, Uppsala University, Norbyvägen 18D, SE-75236 Uppsala, Sweden

Abstract

Sometimes, science takes a big leap forward. This is often due to new technology that allows the study of questions previously difficult or even impossible to address. An example of this is provided in this issue (Vera *et al.* 2008) by the first large-scale attempt toward genome sequencing of an ecologically important model, based on the new '454-sequencing technology'. Using this new technology, the protein-coding sequences of the Glanville fritillary butterfly genome have now been largely characterized.

Keywords: EST sequencing, Glanville fritillary butterfly, ultra-high throughput

Received 8 December 2007; revision accepted 11 January 2008

What are the benefits of all this information?

1 It gives direct access to sequence information necessary for candidate-gene approaches or for identification of candidate genes in regions showing signals of linkage to phenotypic traits.

2 Since pools of individuals were used for sequencing and given the depth of coverage obtained, a large number of sequence variants were detected. About seven segregating sites per 1000 bp (note that this rate is highly dependent on the number of individuals screened) were seen using stringent criteria for polymorphism detection. This will provide a valuable resource of genetic markers for QTL, or association mapping and population genetic analysis.

3 With the availability of large-scale sequence information from the transcriptome (the part of the genome that is transcribed into RNA), species-specific microarray can be constructed. Vera and colleagues (2008) constructed an oligonucleotide array (60 mers) and performed a small expression pilot study, providing proof-of-principle for the feasibility of the approach.

Admittedly, the new study contains limited ecological information. It also ignores the potential for molecular evolutionary analysis using a comparative genomic approach in which sequences of *M. cinxia* would be aligned to orthologs in *B. mori*, *Drosophila melanogaster* and other insects, and surveyed for signs of selection. The strength of the study lies in that it offers a resource for further analyses of adaptation and population processes in an ecologically important model; analyses that otherwise would have been more or less impossible to do. A specific example of this was the surprising discovery of over 400 genes from the intracellular fungal parasite microsporidia, which where subsequently added to the microarray. Microsporidia are commonly found in insects and are known to affect insect physiology and population-level dynamics. More generally, it may be expected that this study should stimulate others to take the step into large-scale genomics, paving the way for a new generation of molecular ecological

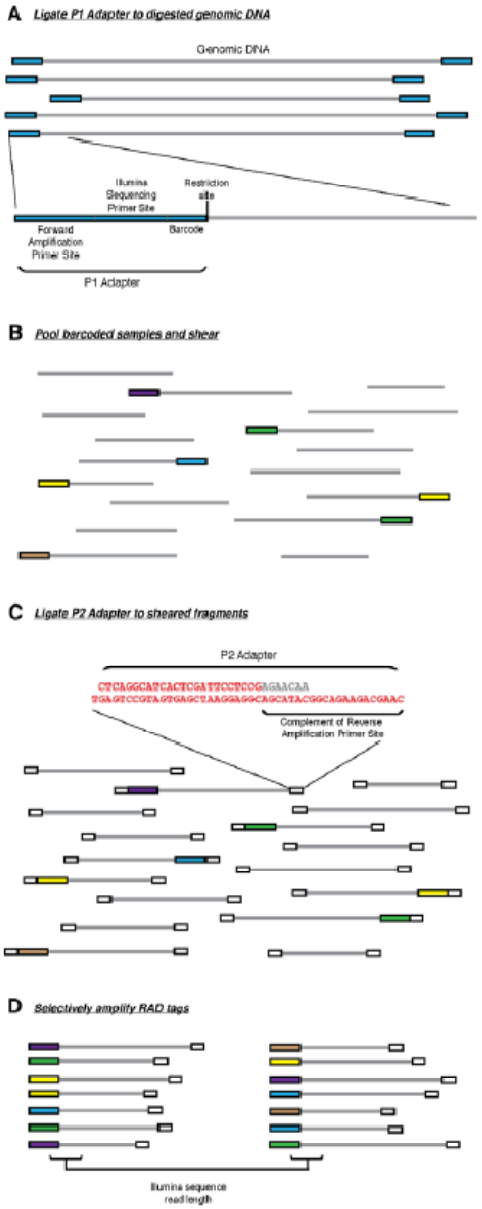


Figure 1. RAD marker generation. (A) Genomic DNA was digested with a restriction enzyme and the P1 adapter was ligated to the fragments. The P1 adapter contains a forward amplification primer site, an Illumina sequencing primer site, and a barcode (colored boxes represent P1 adapters with different barcodes). (B) Adapter-ligated fragments were combined (if multiplexing), sheared and (C) ligated to a second adapter (P2, white boxes). The P2 adapter is a divergent "Y" adapter, containing the reverse complement of the reverse amplification primer site preventing amplification of genomic fragments lacking a P1 adapter. (D) RAD tags, which have a P1 adapter, will be selectively and robustly enriched.
doi:10.1371/journal.pone.0003376.g001

OPEN ACCESS Freely available online PLoS on

Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers

Nathan A. Baird[1], Paul D. Etter[1], Tressa S. Atwood[2], Mark C. Currey[3], Anthony L. Shiver[1], Zachary A Lewis[1], Eric U. Selker[1], William A. Cresko[3], Eric A. Johnson[1]*

1 Institute of Molecular Biology, University of Oregon, Eugene, Oregon, United States of America, **2** Floragenex, Eugene, Oregon, United States of America, **3** The Cent for Ecology and Evolutionary Biology, University of Oregon, Eugene, Oregon, United States of America

Restriction-site Associated DNA (RAD) Tags

Abstract

Single nucleotide polymorphism (SNP) discovery and genotyping are essential to genetic mapping. There remains a need for a simple, inexpensive platform that allows high-density SNP discovery and genotyping in large populations. Here we describe the sequencing of restriction-site associated DNA (RAD) tags, which identified more than 13,000 SNPs, and mapped three traits in two model organisms, using less than half the capacity of one Illumina sequencing run. We demonstrated that different marker densities can be attained by choice of restriction enzyme. Furthermore, we developed a barcoding system for sample multiplexing and fine mapped the genetic basis of lateral plate armor loss in threespine stickleback by identifying recombinant breakpoints in F_2 individuals. Barcoding also facilitated mapping of a second trait, a reduction of pelvic structure, by *in silico* re-sorting of individuals. To further demonstrate the ease of the RAD sequencing approach we identified polymorphic markers and mapped an induced mutation in *Neurospora crassa*. Sequencing of RAD markers is an integrated platform for SNP discovery and genotyping. This approach should be widely applicable to genetic mapping in a variety of organisms.

120 chromosomes

250 chromosomes

370 chromosomes

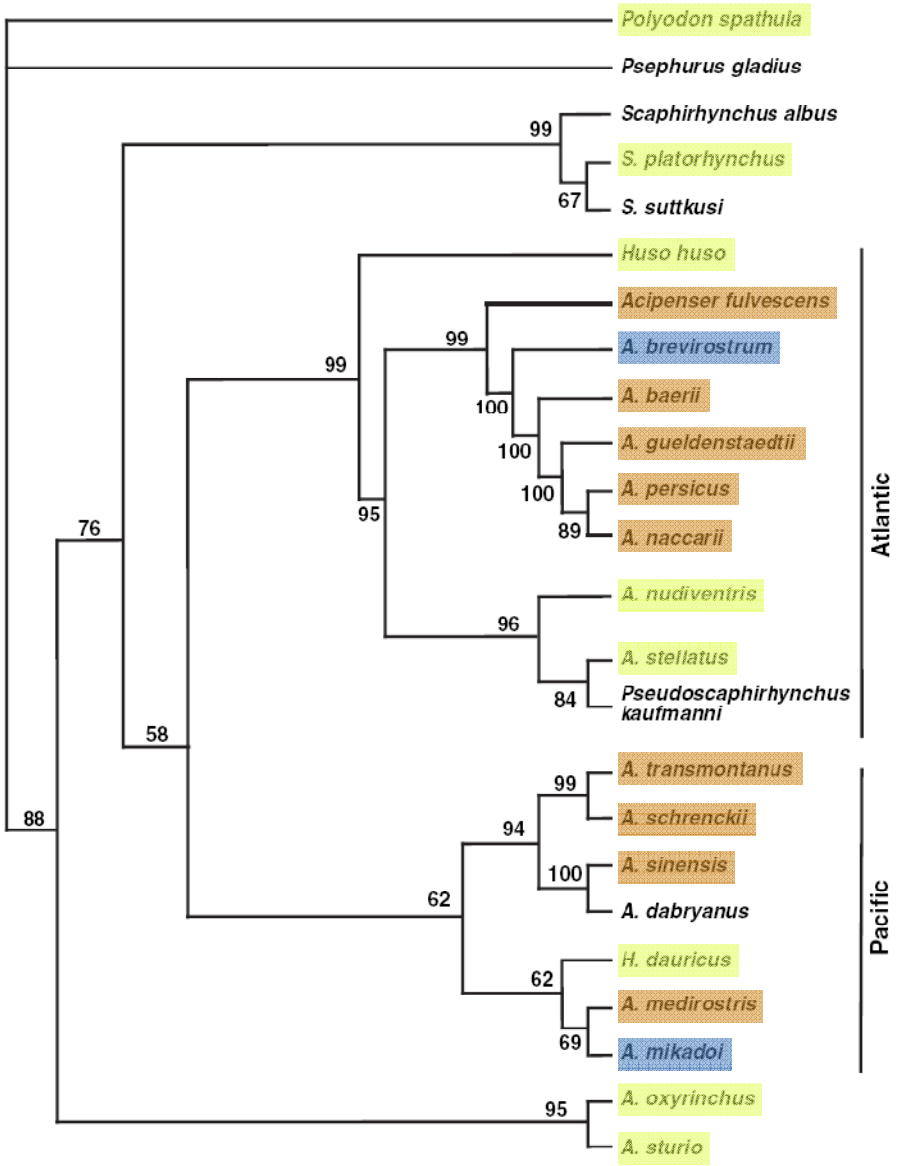
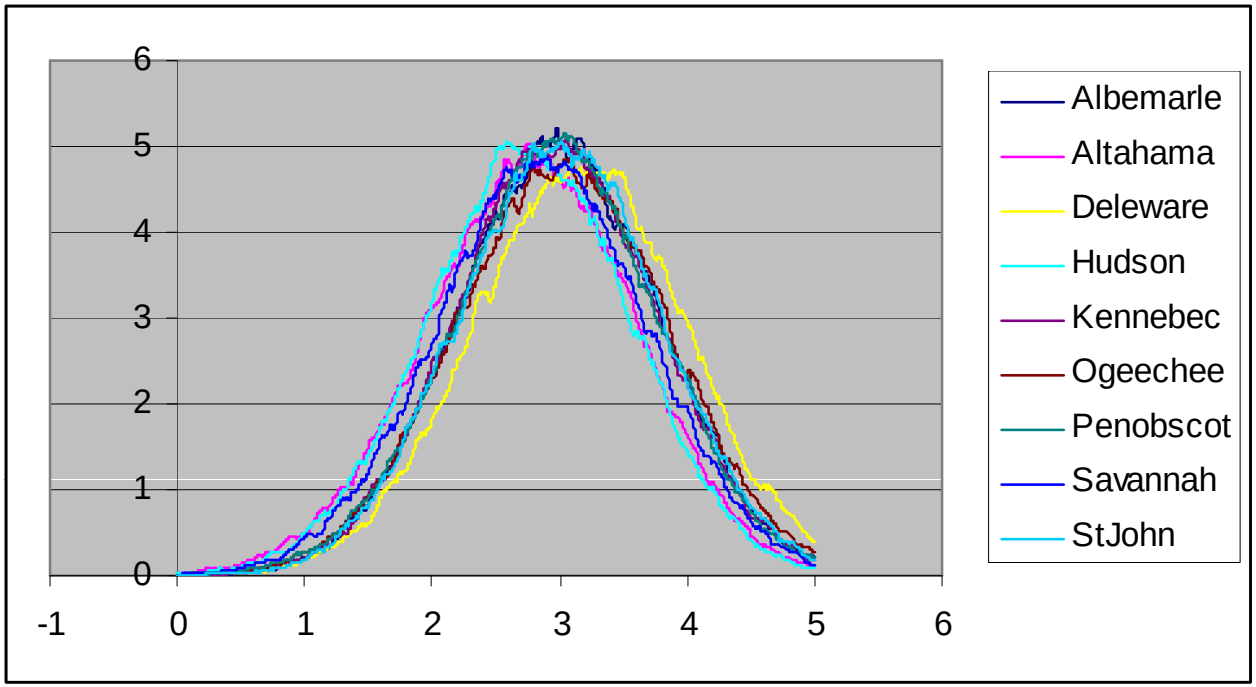


Fig. 1. Phylogenetic tree based on combined mitochondrial cytochrome *b*, 12S rRNA, cytochrome *c* oxidase subunit II, $tRNA_{Phe}$, $tRNA_{Asp}$, 16S rRNA, NADH5 and control region gene sequences (eight genes, 4406 sites), excluding *Acipenser ruthenus*, produced using a maximum likelihood approach with 10 000 puzzling steps (log $L = -17694.14$). The numbers located above each branch indicate the quartet puzzling values (as a percent) of the node. Trees produced with Bayesian, neighbor-joining distance and maximum parsimony methods (not shown) were similar in topology (see text)

Unique Demographic Perspective



Posterior distribution of Log(T)

1. Little indications of any difference in T between populations.
2. Old decline in populations with continuing gene flow.
3. A decline in the central range which over time have lead to a decline in Ne in adjacent populations as the genetic variation in the incoming migrants was gradually lowered.
4. Concordant with shallow mtDNA gene genealogy.