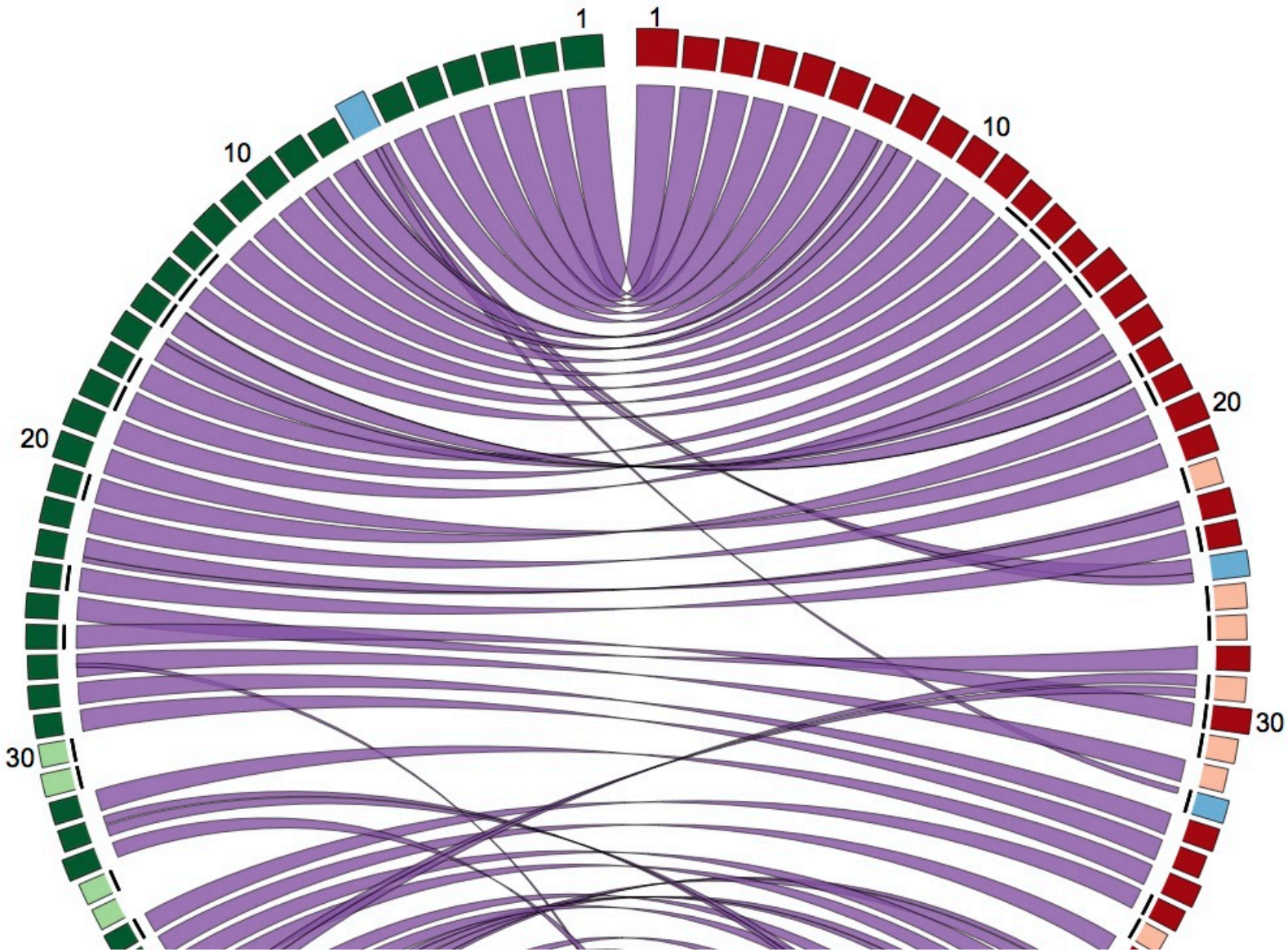
Comparative Genomics



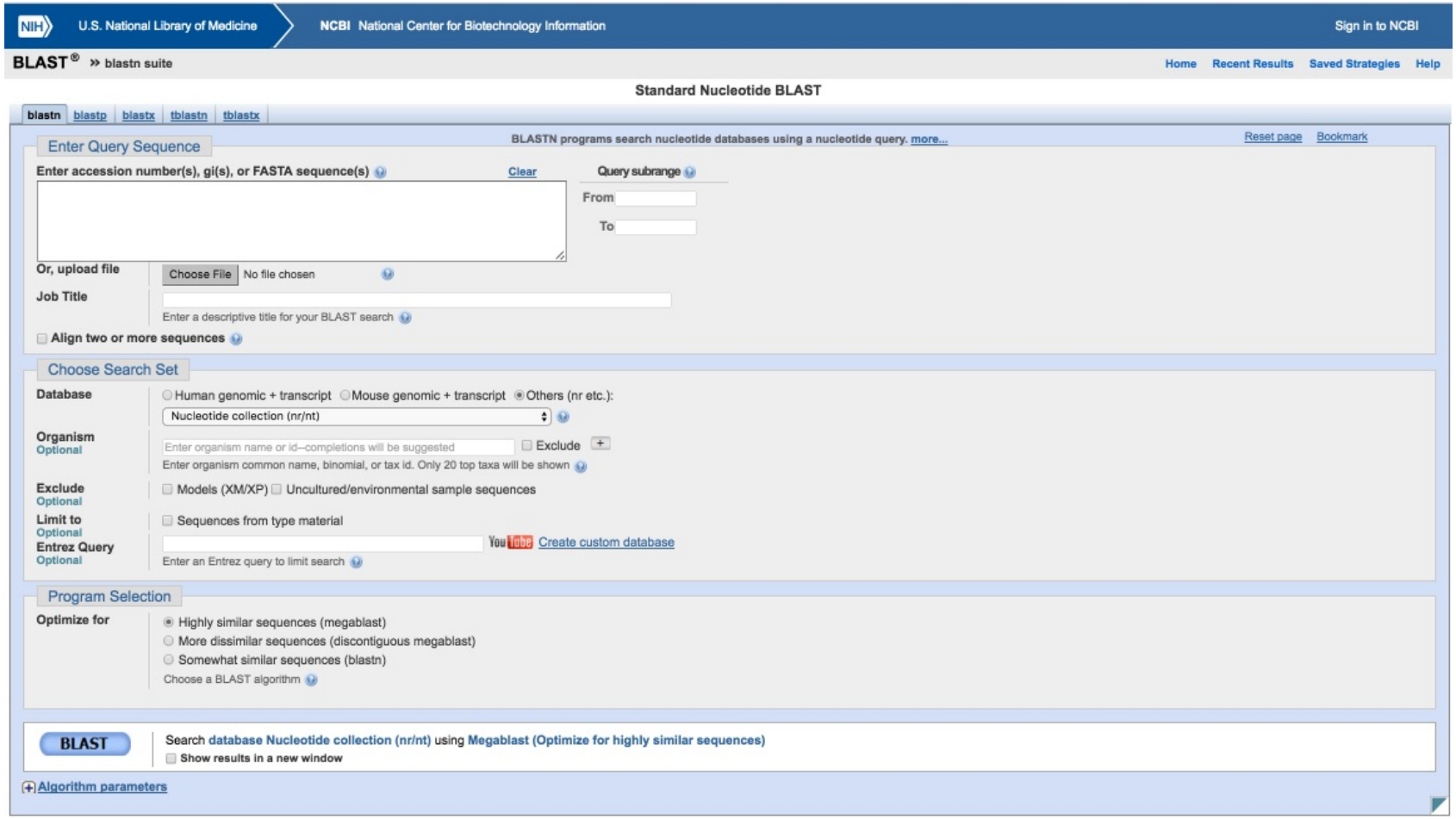
Comparative Genomics

Common Themes

- Gene and functional pathway presence/absence
- Identification, alignment of orthologs
- Structural analysis and comparison

Comparative Genomics

Basic Local Alignment Search Tool (BLAST)



Comparative Genomics

Colorado State University

The Core BLAST Programs

Program	Query Sequences	Database Sequences	Alignment
blastn	nucleotide	nucleotide	nucleotide
blastp	amino acid	amino acid	amino acid
blastx	nucleotide	amino acid	amino acid
tblastn	amino acid	nucleotide	amino acid
tblastx	nucleotide	nucleotide	amino acid

Comparative Genomics

Downstream Applications

- Identify orthologs
  - Reciprocal BLAST
  - OrthoMCL
- Annotation and Pathway Analysis
  - e.g., BLAST2GO
- Many others

Comparative Genomics

Colorado State University

Advantages of Running BLAST Locally

- Customize databases with unpublished sequences
- Run thousands of simultaneous searches
- Integrate into custom scripts and pipelines

Comparative Genomics

Parsing BLAST Output

```
BLASTN 2.2.30+

Database: bacterial-tRNAs.fas
           24,746 sequences; 1,923,058 total letters

Query= 1

Length=74
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  Microcystis_aeruginosa_NIES_843_chr.trna42-AspGTC (409004-40893...    111   2e-26
  Anabaena_variabilis_ATCC_29413_chr.trna22-AspGTC (6126075-61260...    111   2e-26

> Microcystis_aeruginosa_NIES_843_chr.trna42-AspGTC (409004-408931)
 Asp (GTC) 74 bp  Sc: 75.49
Length=74

 Score =   111 bits (122),  Expect = 2e-26
 Identities = 69/74 (93%), Gaps = 0/74 (0%)
 Strand=Plus/Plus

Query  1   GGGATTGTAGTTCAATTGGTTAGAGCACCGCCCTGTCAAGGCGGAAGCTACGGGTTCGAG  60
           ||||||||||||||||||||||||||||||||||||| |||||||| | ||||||||||
Sbjct  1   GGGATTGTAGTTCAATTGGTTAGAGCACCGCCCTGTCACGGCGGAAGTTGCGGGTTCGAG  60

Query  61  TCCCGTCAGTCCCG  74
           ||||||| |||||
Sbjct  61  CCCCGTCAATCCCG  74
```

Comparative Genomics

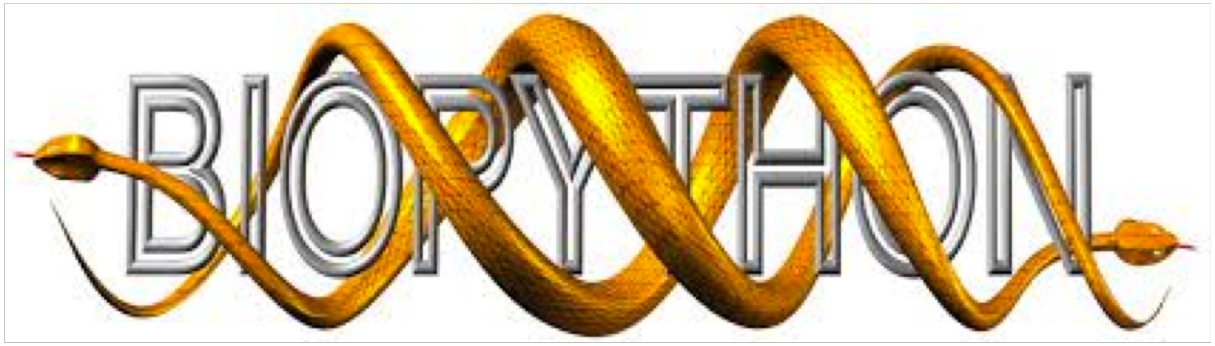
Colorado State University

Organization of BLAST Output

- RESULTS – One results block for each sequence in the original query file
    - HITS – Within each results block, one hit block for each sequence in the database that produced a significant alignment
        - HSPs – Within each hit block, one high-scoring-pair block for each significant local alignment to that hit

Comparative Genomics

Parsing BLAST Output



http://bioperl.org/howtos/SearchIO_HOWTO.html

Comparative Genomics

Parsing BLAST Output – BioPerl Results Information

Object	Method	Example	Description
Result	algorithm	BLASTX	algorithm string
Result	algorithm_version	2.2.4 [Aug-26-2002]	algorithm version
Result	query_name	gi	20521485
Result	query_accession	AP004641.2	query accession
Result	query_length	3059	query length
Result	query_description	Oryza sativa ... 977CE9AF checksum.	query description
Result	database_name	test.fa	database name
Result	database_letters	1291	number of residues in database
Result	database_entries	5	number of database entries
Result	available_statistics	effectivespaceused ... dbletters	statistics used
Result	available_parameters	gapext matrix allowgaps gapopen	parameters used
Result	num_hits	1	number of hits
Result	hits		List of all Bio::Search::Hit::GenericHit objects for this Result
Result	rewind		Reset the internal iterator that dictates where next_hit() is pointing, useful for re-iterating through the list of hits

Comparative Genomics

Parsing BLAST Output – BioPerl Hit Information

Object	Method	Example	Description
Hit	name	gb	443893
Hit	length	331	Length of the Hit sequence
Hit	accession	443893	accession (usually when this is a Genbank formatted id this will be an accession number - the part after the `gb` or `emb`)
Hit	description	LaForas sequence	hit description
Hit	algorithm	BLASTX	algorithm
Hit	raw_score	92	hit raw score
Hit	significance	2e-022	hit significance
Hit	bits	92.0	hit bits
Hit	hsps		List of all `Bio::Search::HSP::GenericHSP` objects for this Hit
Hit	num_hsps	1	number of HSPs in hit
Hit	locus	124775	locus name
Hit	accession_number	443893	accession number
Hit	rewind		Resets the internal counter for next_hsp() so that the iterator will begin at the beginning of the list

Comparative Genomics

Parsing BLAST Output – BioPerl HSP Information

Object	Method	Example	Description
HSP	algorithm	BLASTX	algorithm
HSP	evalue	2e-022	e-value
HSP	expect	2e-022	alias for evalue()
HSP	frac_identical	0.884615384615385	fraction identical
HSP	frac_conserved	0.923076923076923	fraction conserved (conservative and identical replacements aka "fraction similar")
HSP	gaps	2	number of gaps
HSP	query_string	DMGRCSSG ..	query string from alignment
HSP	hit_string	DIVQNSS ...	hit string from alignment
HSPt	homology_string	D+ + SSGCN ...	string from alignment
HSP	length('total')	52	length of HSP (including gaps)
HSP	length('hit')	50	length of hit participating in alignment minus gaps
HSP	length('query')t	156	length of query participating in alignment minus gaps
HSPt	hsp_length	52	Length of the HSP (including gaps) alias for length('total')
HSPt	frame	0	$hsp->query->frame,$hsp->hit->frame
HSP	num_conserved	48	number of conserved (conservative replacements, aka "similar") residues
HSP	num_identical	46	number of identical residues
HSPt	rank	1	rank of HSP
HSP	seq_inds('query','identical')	(966,971,972,973,974,975 ...)	identical positions as array
HSP	seq_inds('query','conserved-not-identical')	(967,969)	conserved, but not identical positions as array
HSP	seq_inds('query','conserved')	(966,967,969,971,973,974,975, ...)	conserved or identical positions as array
HSP	seq_inds('hit','identical')	(197,202,203,204,205, ...)	identical positions as array
HSP	seq_inds('hit','conserved-not-identical')	(198,200)	conserved not identical positions as array
HSP	seq_inds('hit','conserved',1)	(197,202-246)	conserved or identical positions as array, with runs of consecutive numbers compressed

Comparative Genomics

Parsing BLAST Output – BioPerl HSP Information (Cont.)

HSPt	score	227	score
HSP	bits	92.0	score in bits
HSP	range('query')	(2896,3051)	start and end as array
HSP	range('hit')	(197,246)	start and end as array
HSP	percent_identity	88.4615384615385	% identical
HSP	strand('hit')	1	strand of the hit
HSP	strand('query')	1	strand of the query
HSP	start('query')	2896	start position from alignment
HSP	end('query')	3051	end position from alignment
HSP	start('hit')	197	start position from alignment
HSP	end('hit')	246	end position from alignment
HSP	matches('hit')	(46,48)	number of identical and conserved as array
HSP	matches('query')	(46,48)	number of identical and conserved as array
HSP	get_aln	sequence alignment	Bio::SimpleAlign object
HSPt	hsp_group	*Not available in this report*	Group field from WU-BLAST reports run with -topcomboN or -topcomboE specified
HSP	links	*Not available in this report*	Links field from WU-BLAST reports run with -links showing consistent HSP linking

Exercise

https://dbsloan.github.io/TS2019/exercises/local_blast.html

- Make BLAST databases
- Run local BLAST searches
- Parse BLAST output with BioPerl
- Make dot-plot comparing two bacterial genomes in R using parsed BLAST output