



Bioinformatics Analysis Tools and the PoPLAR Science Gateway

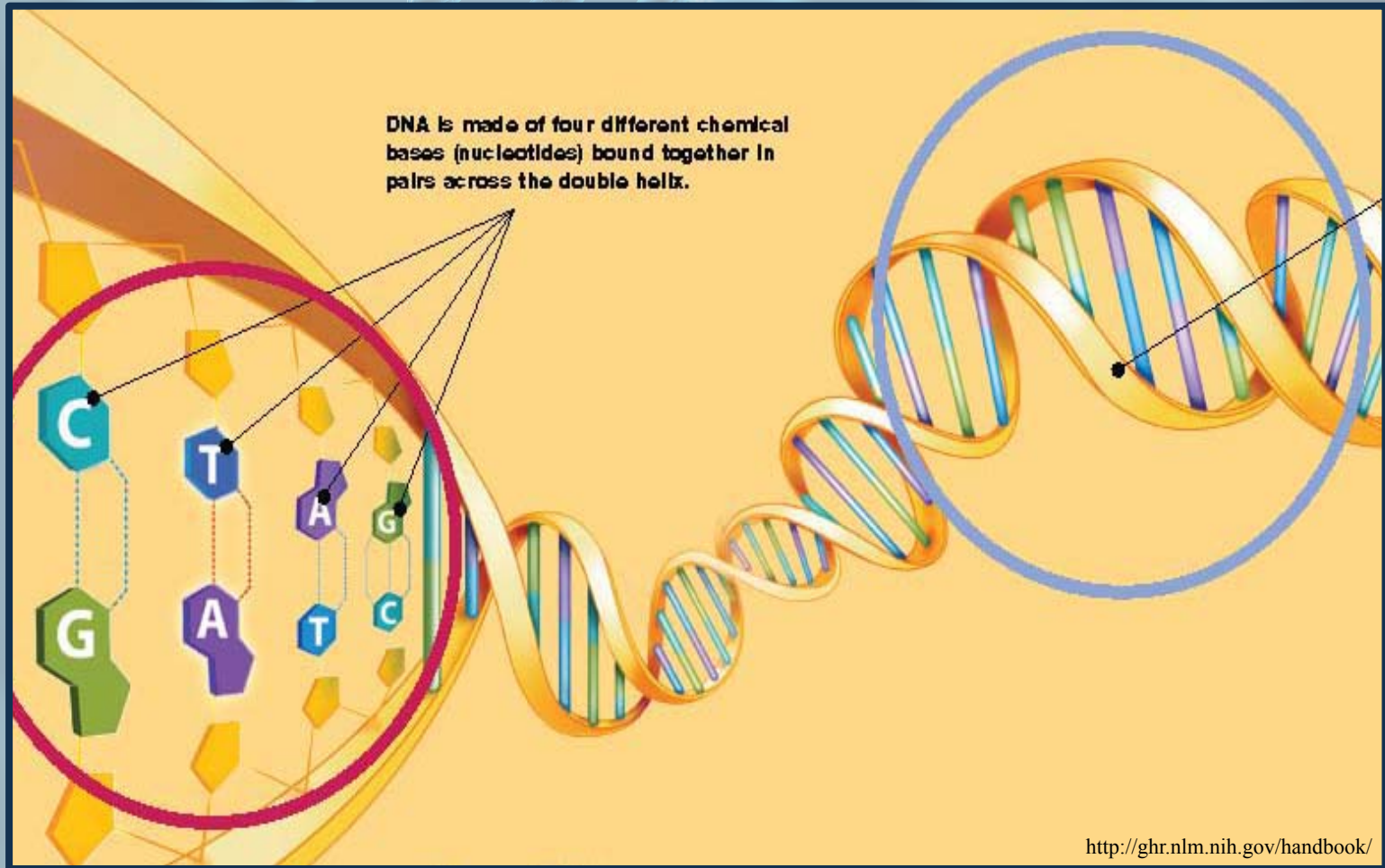
Mary Lauren Harris (Baylor University)
Mentor: Bhanu Rekepalli, PhD (NICS)

Highlights

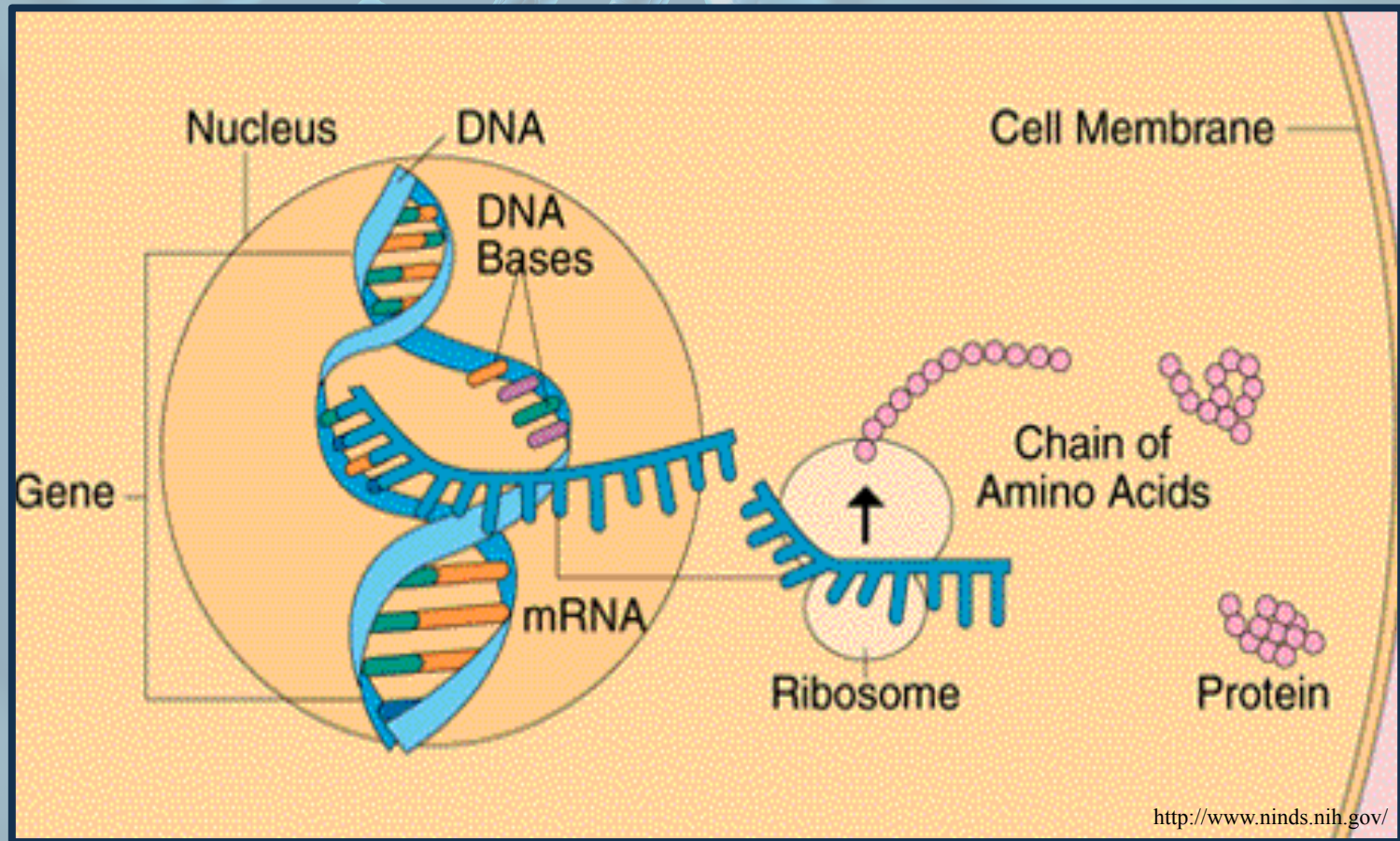


- Background
- Challenges
- Genome Analysis Tools
- HSP-BLAST
- Results
- Science Gateway
- Future Work

Genes



Proteins



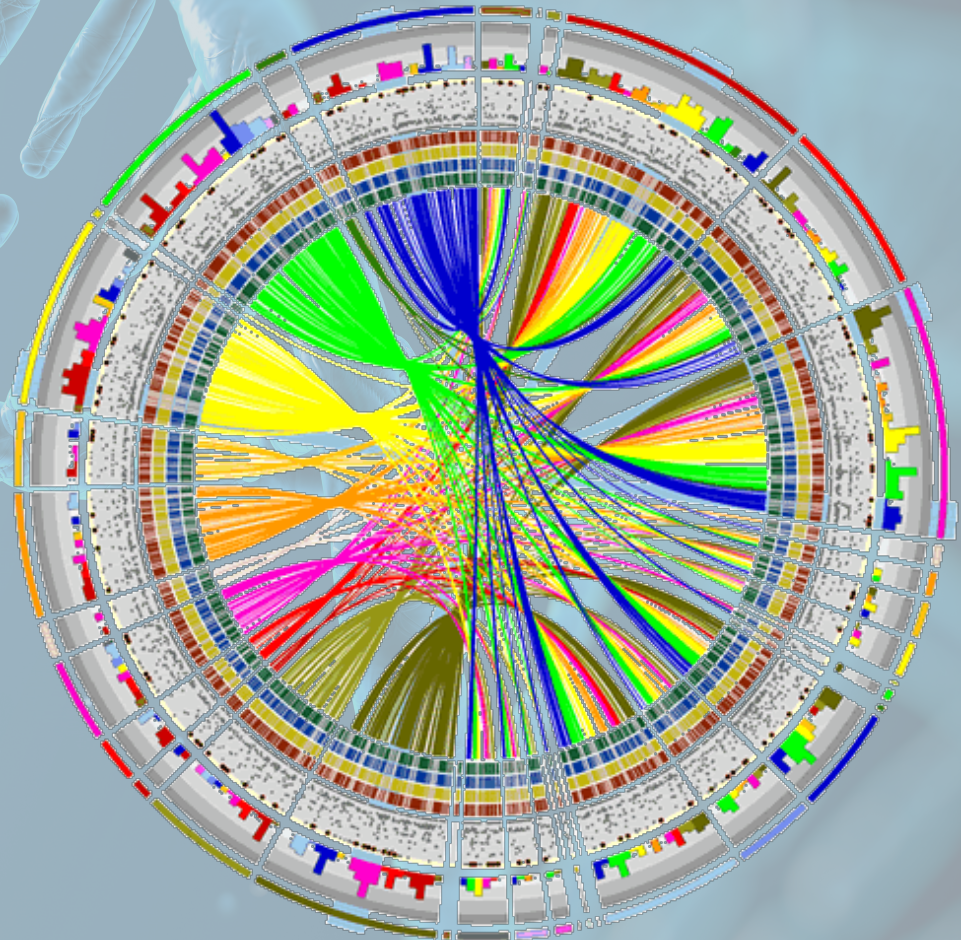
Bioinformatics

Analyze

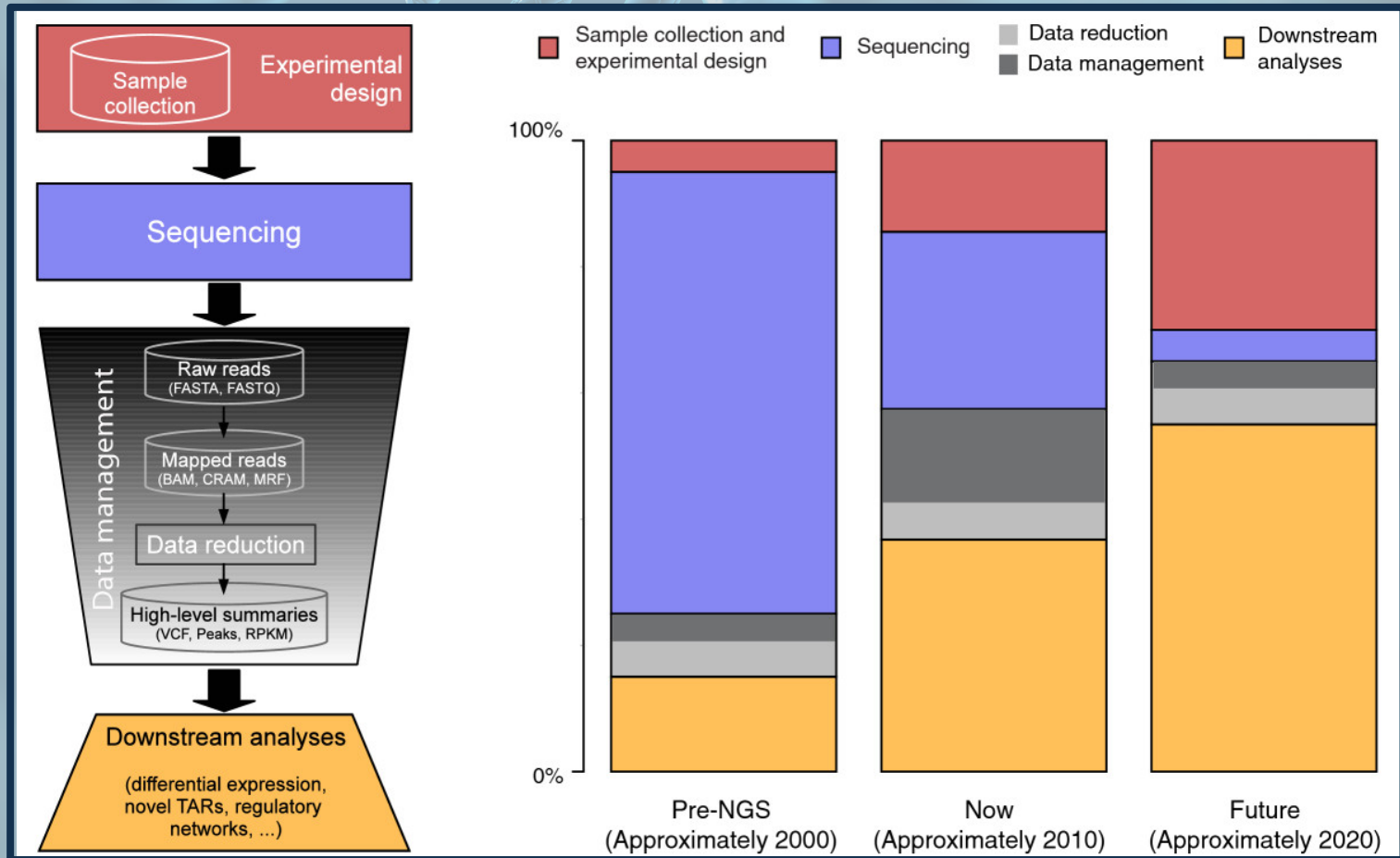
Manage

Discover

Visualize

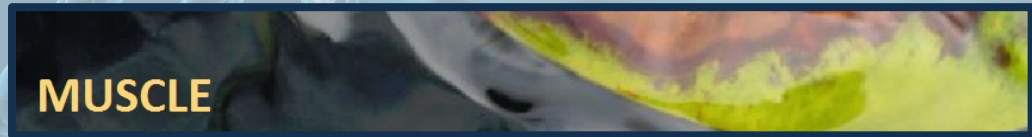


Challenges



Analysis Tools

MUSCLE

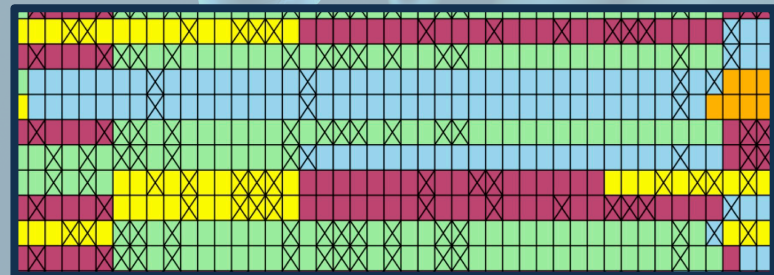


HMMER



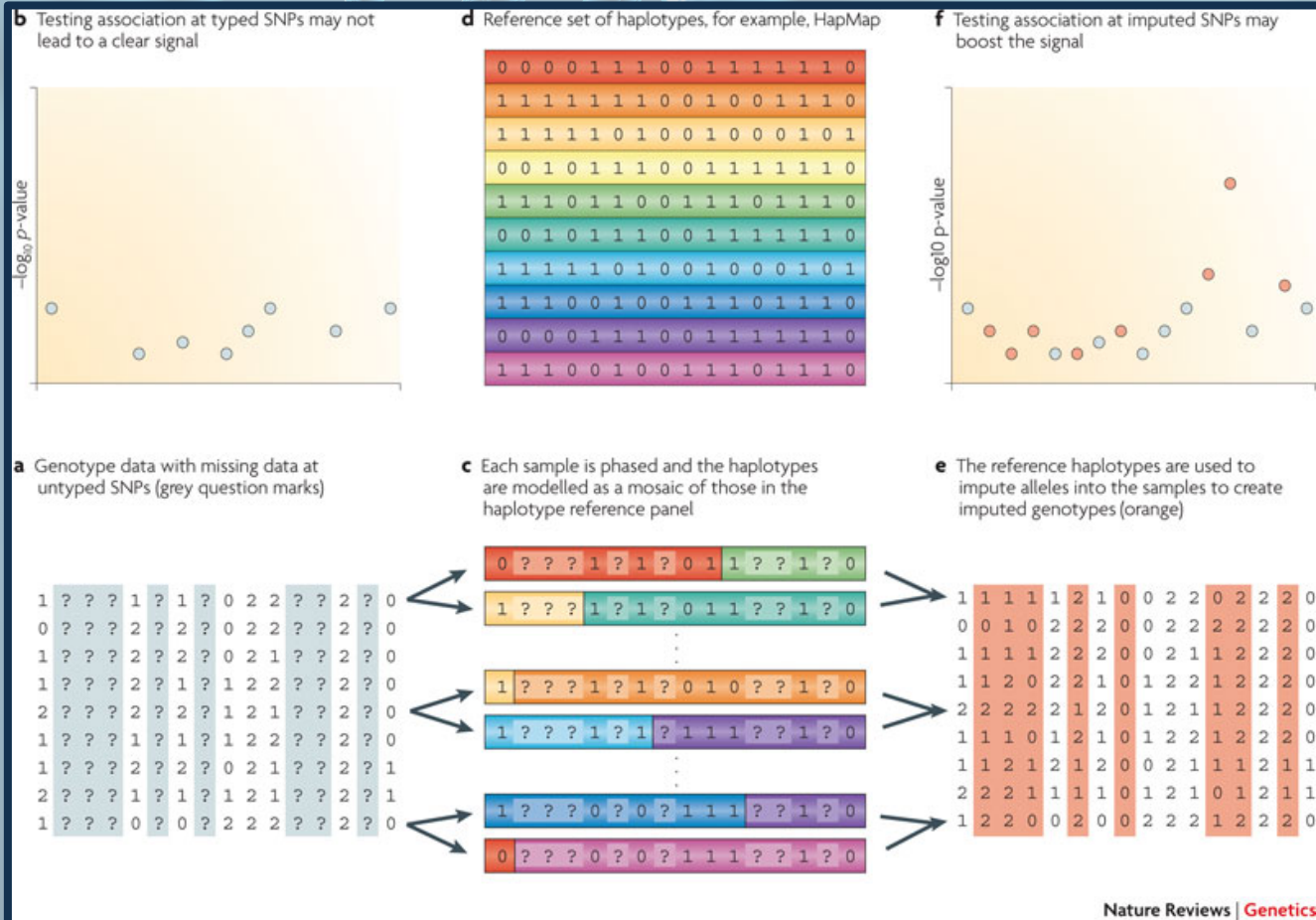
BLAST

MaCH



Minimac

Genotype Imputation



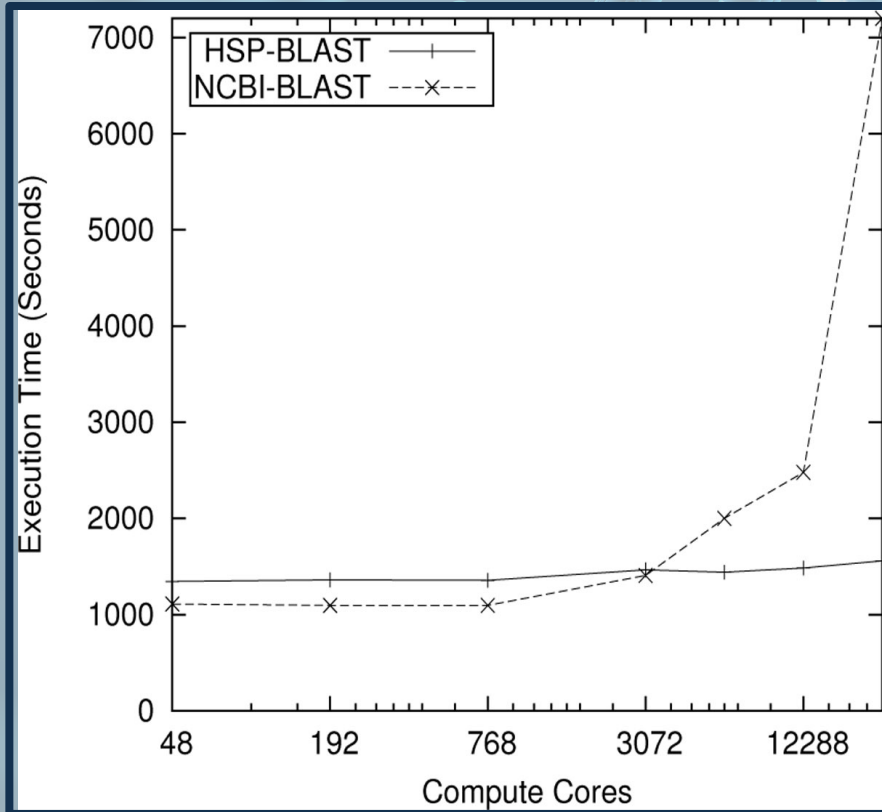
Howie, B. Genotype imputation for genome-wide association studies.

MaCH Documentation

```
1 MaCH is a tool for haplotyping, genotype imputation and disease association analysis
2 developed by Goncalo Abecasis and Yun Li. MaCH was first used to imputed missing
3 genotypes in our FUSION genome-wide association study (Scott et al, Science, 2007)
4 and has since been used in the analysis of many other GWAS.
5
6 This file explains how to install MaCH 1.0 on NICS computing resources, although
7 the instructions are relevant for all Linux systems. Below the numbered instructions
8 is a modified and extended version of the README file provided by the developers
9 to explain the options, inputs, and how to run MaCH 1.0. This file includes
10 information from both the MaCH homepage and wiki as well.
11
12 Contents:  Installation, Input, Options, Execution/Examples, Output
13             Quality Assessment, Further Analysis, Troubleshooting,
14             Additional Resources
15
16 =====
17 = INSTALL MACH 1.0 =
18 =====
19
20 1. Get the tar file from the MaCH website using the following
21    command:
22
23     wget http://www.sph.umich.edu/csg/abecasis/MaCH/download/mach.1.0.18.source.tgz
24
25 2. Untar the files using the tar command.
26
27     tar -xf mach.1.0.18.source.tgz
28
29 3. Navigate to the directory containing 'makefile,' if
30    necessary.
31
32     cd directory/with/new/files
33
34 4. Ensure you are using g++ as the compiler. If you use NICS
35    resources, you should use the CC wrapper; and you may have
36    to swap modules. Use 'module list', 'module avail' and
37    'module swap <old> <new>' commands to accomplish this.
38
39     make all
40
41    OR (NICS-- ie. Darter)
42
43     module swap PrgEnv-cray PrgEnv-gnu
44     module list (look for gcc)
45
46     make all CXX-CC (to use the CC compiler wrapper)
47
48 5. Install the binaries in /usr/local/bin by using the
49    following command:
50
51     make install
52
53    OR
54
55     make install=/directory/to/install
```

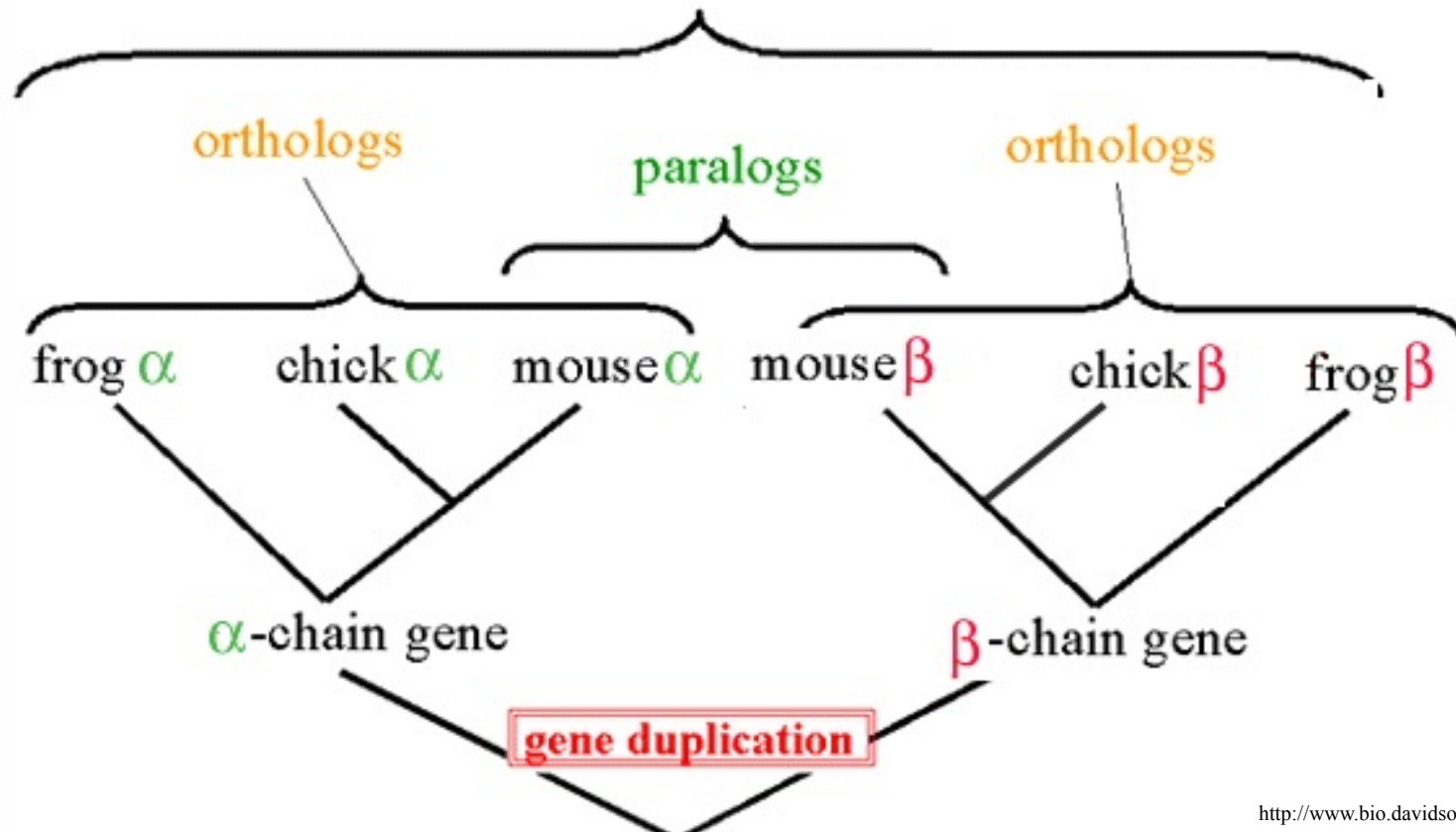
```
5 = STRATEGY #2 =
6 USE REFERENCE (eg. HAPMAP) HAPLOTYPES AS INPUT:
7
8 FILES:
9
10 If you select this option, you should generate a file that includes a set of
11 reference haplotypes. These can be typed at more markers than are available
12 in your sample. You will also need a small file that lists all the markers
13 that appear in the phased haplotypes.
14
15 Then, to estimate missing genotypes, you'll need to provide the Merlin format
16 data and pedigree files, the reference haplotypes and the list of SNPs in the
17 reference haplotypes.
18
19 USEFUL COMMAND LINE OPTIONS:
20
21 Name the reference haplotype and snp list files with --haps and --srps.
22
23 If you use the --autoFlip option, MaCH 1.0 will try to automatically resolve
24 problems with alleles that are inconsistently labeled in your sample and the
25 reference panel (by flipping strands and dropping markers where this trivial
26 solution does not help).
27
28 Most of the time, you'll get good estimates of genotypes at untyped markers
29 using the --rounds <k> and --greedy option.
30
31 If you don't use the --greedy option, you can control computational effort
32 with the --weighted and --states <k> options. However, this alternative strategy
33 generally requires more iterations before converging to a good solution.
34
35 EXAMPLE USAGE:
36
37 mach -d sample.dat -p sample.ped -h hapmap.haplos -s hapmap.srps --rounds 50 --greedy --geno
38
39 mach -d sample.dat -p sample.ped -h hapmap.haplos -s hapmap.srps --rounds 500 --states 200 --geno
40
41 mach -d sample.dat -p sample.ped -h hapmap.haplos -s hapmap.srps --rounds 500 --states 200 --weighted --geno
42
43 NOTE: It is very important to ensure that alleles are labeled consistently in
44 your sample and in the reference panel. MaCH 1.0 will automatically warn
45 you about alleles that differ in frequency greatly between your sample
46 and the reference panel or that have different allele names in the two
47 subsets of data. However, these checks will not catch all inconsistently
48 labeled alleles.
49
50 = STRATEGY #3 =
51 USING DATA FROM PREVIOUS RUNS TO SPEED UP IMPUTATION
52
53 The standard genotype imputation approach, described as strategy #2 works best
54 when you execute a large number of iterations of the Markov Chain (50-100). These
55 iterations are used to simultaneously update the crossover map (which determines
56 the likely locations for haplotype transitions), to update the errorrate map
57 (which flags unusual markers), and to estimate the missing genotypes.
58
59 An alternative approach is to use a single set of estimates for the crossover and
```

HSP-BLAST

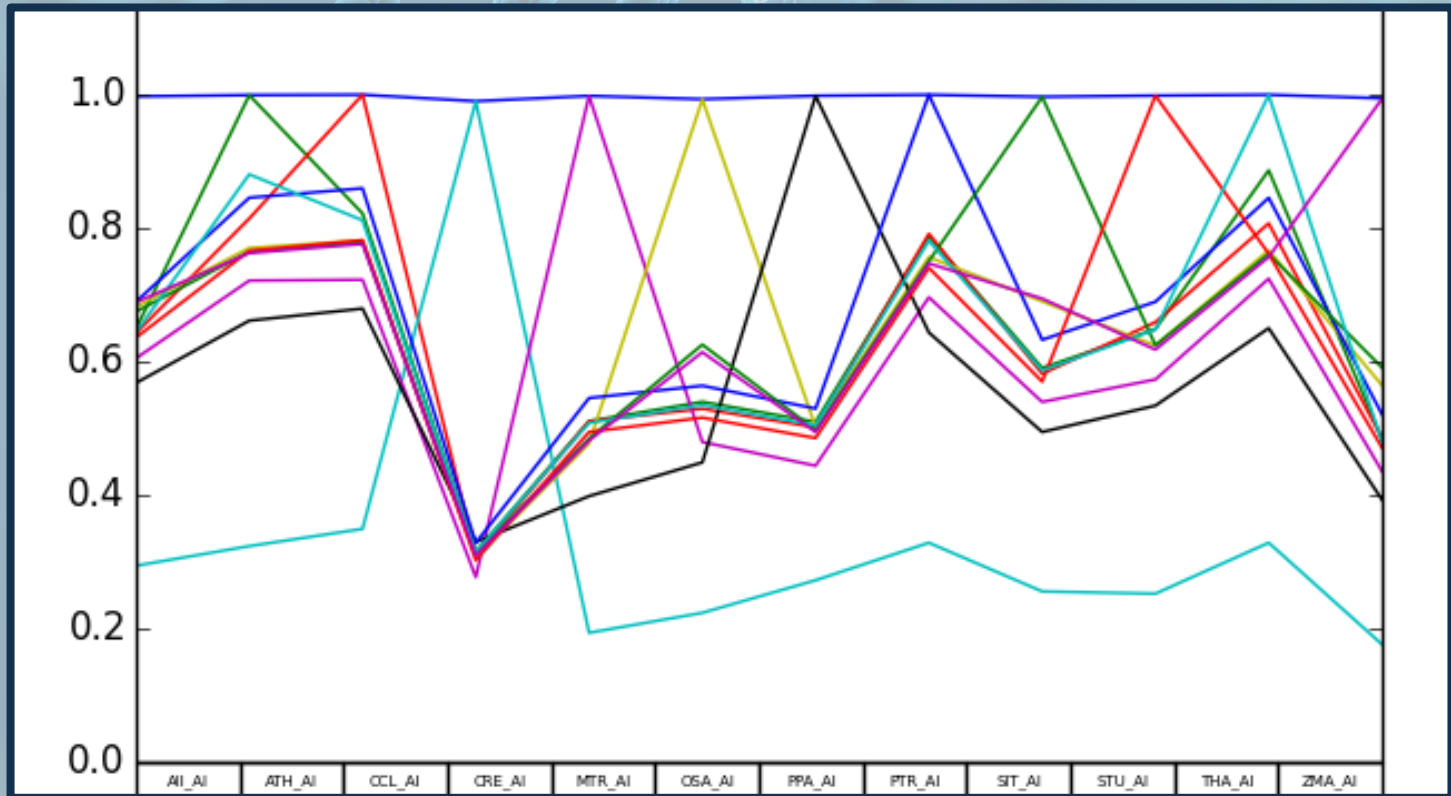


ATH_AT4G10265.1	PTR_0190117200.1	63.64	88	25	2	1	82	1	87	1e-18	69.0
ATH_AT4G10265.1	PTR_0130148100.1	61.80	89	26	2	1	82	1	88	2e-18	88.6
ATH_AT4G10265.1	PTR_0190116500.1	60.67	89	27	2	1	82	1	88	3e-18	87.4
ATH_AT4G10265.1	STU_32044	59.77	87	29	2	1	82	1	86	5e-18	87.0
ATH_AT4G10265.1	SIT_12094a	55.42	83	36	1	1	82	1	83	1e-17	85.9
ATH_AT4G10265.1	PTR_1163500.1	61.36	88	27	2	1	82	1	87	2e-17	85.1
ATH_AT4G10265.1	PTR_0190117200.2	61.36	88	27	2	1	82	1	88	4e-17	85.1
ATH_AT4G10265.1	PTR_0190116800.1	55.06	89	32	2	1	82	1	85	4e-17	84.0
ATH_AT4G10265.1	PTR_0190117700.1	60.47	86	29	2	1	82	1	87	4e-17	84.0
ATH_AT4G10265.1	ZHA_GRMZ020106393	50.57	87	38	2	1	82	1	87	8e-17	83.2
ATH_AT4G10265.1	PTR_0130148000.1	53.93	89	33	2	1	82	1	88	1e-16	82.4
ATH_AT4G10265.1	SIT_11483m	51.19	84	39	2	1	82	1	84	2e-16	81.6
ATH_AT4G10265.1	CCL_47345m	56.82	88	28	3	1	83	1	83	1e-15	79.8
ATH_AT4G10265.1	ZHA_GRMZ020106413	52.44	82	33	2	7	82	11	92	2e-15	78.6
ATH_AT4G10265.1	ZHA_GRMZ020106445	48.19	83	38	2	1	82	1	79	5e-15	77.0
ATH_AT4G10265.1	CCL_17327a	57.30	89	30	3	1	83	1	87	5e-15	77.0
ATH_AT4G10265.1	PTR_0010408300.1	51.06	94	33	3	2	83	3	95	3e-13	71.2
ATH_AT4G10265.1	PTR_0010408400.1	45.65	92	41	2	1	83	1	92	5e-13	70.5
ATH_AT4G10265.1	SIT_11449a	45.00	80	35	1	13	83	18	97	4e-12	67.4
ATH_AT4G10265.1	PTR_0190116600.1	59.55	89	28	2	1	82	1	88	6e-12	66.6
ATH_AT4G10265.1	PTR_0130148200.1	54.32	81	32	2	6	82	1	80	1e-11	66.2
ATH_AT4G10265.1	CCL_17300a	52.63	95	32	2	1	83	1	94	3e-11	64.7
ATH_AT4G10265.1	STU_32043	54.12	85	27	3	1	82	1	76	4e-11	63.9
ATH_AT4G10265.1	CCL_47315m	60.00	90	27	2	1	82	1	89	4e-11	63.9
ATH_AT4G10265.1	OSA_LOD_0s00g08090.1	45.83	72	38	1	13	83	19	90	3e-10	60.8
ATH_AT4G10265.1	CCL_17642a	53.57	84	33	2	1	79	1	83	1e-09	58.9
ATH_AT4G10265.1	CCL_17290a	50.00	95	34	2	1	83	1	95	2e-09	58.2
ATH_AT4G10265.1	OSA_LOD_0s04g04230.1	49.84	79	37	1	13	83	18	96	9e-09	55.2
ATH_AT4G10265.1	OSA_LOD_0s02g06560.1	42.35	85	39	3	3	82	0	87	2e-08	55.1
ATH_AT4G10265.1	SIT_08830a	34.83	89	51	1	1	82	6	94	2e-07	51.6
ATH_AT4G10265.1	ZHA_GRMZ020419994	61.76	34	13	0	2	35	6	39	2e-06	48.5
ATH_AT4G10265.1	ZHA_AC225718.2_FG	57.14	35	15	0	1	35	7	41	2e-06	48.5
ATH_AT4G10265.1	PTR_0190117800.1	95.24	21	1	0	62	82	70	90	2e-06	48.1
ATH_AT4G10265.1	CCL_48116m	90.48	21	2	0	62	82	59	79	8e-06	46.6
ATH_AT4G10265.1	OSA_LOD_0s06g46970.1	30.00	90	48	1	8	82	13	102	2e-05	45.1
ATH_AT4G10265.1	ATH_AT4G33260.1	34.12	85	47	2	7	83	8	91	4e-05	43.9
ATH_AT4G10265.1	THA_22930a	33.33	87	39	3	0	83	12	90	6e-05	43.5
ATH_AT4G10265.1	PTR_0190116700.1	60.00	25	8	0	58	82	37	61	2e-04	42.0
ATH_AT4G10265.1	ZHA_GRMZ020106511	35.63	87	39	1	13	82	18	104	2e-04	41.6
ATH_AT4G10265.1	PTR_0110941700.1	68.00	25	8	0	58	82	50	74	7e-04	40.0
ATH_AT4G10265.1	THA_29473a	71.43	21	6	0	61	81	70	90	0.001	39.7
ATH_AT4G10265.1	ATH_AT4G095070.1	68.18	22	7	0	61	82	65	86	0.001	39.7
ATH_AT4G10265.1	CCL_47349m	64.29	42	14	1	1	42	1	41	0.022	35.0
ATH_AT4G10265.1	ZHA_GRMZ0206048102	76.47	17	4	0	67	83	60	76	0.10	32.7
ATH_AT4G10265.1	PTR_0110942000.1	59.09	22	9	0	59	80	49	70	0.10	32.7
ATH_AT4G10265.1	THA_26689a	29.76	84	54	2	1	80	1	83	0.27	31.6
ATH_AT4G10265.1	OSA_LOD_0s04g04210.1	78.57	14	3	0	69	82	79	92	0.29	31.2
ATH_AT4G10265.1	SIT_12787a	71.43	14	4	0	69	82	127	140	0.90	29.6
ATH_AT4G10265.1	ATH_AT4G28240.1	76.32	13	3	0	68	80	74	83	2.9	28.1
ATH_AT4G10265.1	SIT_38999a	81.82	44	29	1	1	44	371	413	3.8	27.7
ATH_AT5G37360.1	ATH_AT5G37360.1	100.00	274	0	0	36	309	36	309	2e-140	496
ATH_AT5G37360.1	THA_27857a	85.13	269	37	1	41	309	41	306	1e-114	410
ATH_AT5G37360.1	CCL_21303a	69.20	263	78	1	47	309	48	307	4e-94	342
ATH_AT5G37360.1	CCL_21307a	69.20	263	78	1	47	309	48	307	4e-94	342
ATH_AT5G37360.1	PTR_0190823700.1	65.88	255	85	1	54	308	51	303	3e-87	319

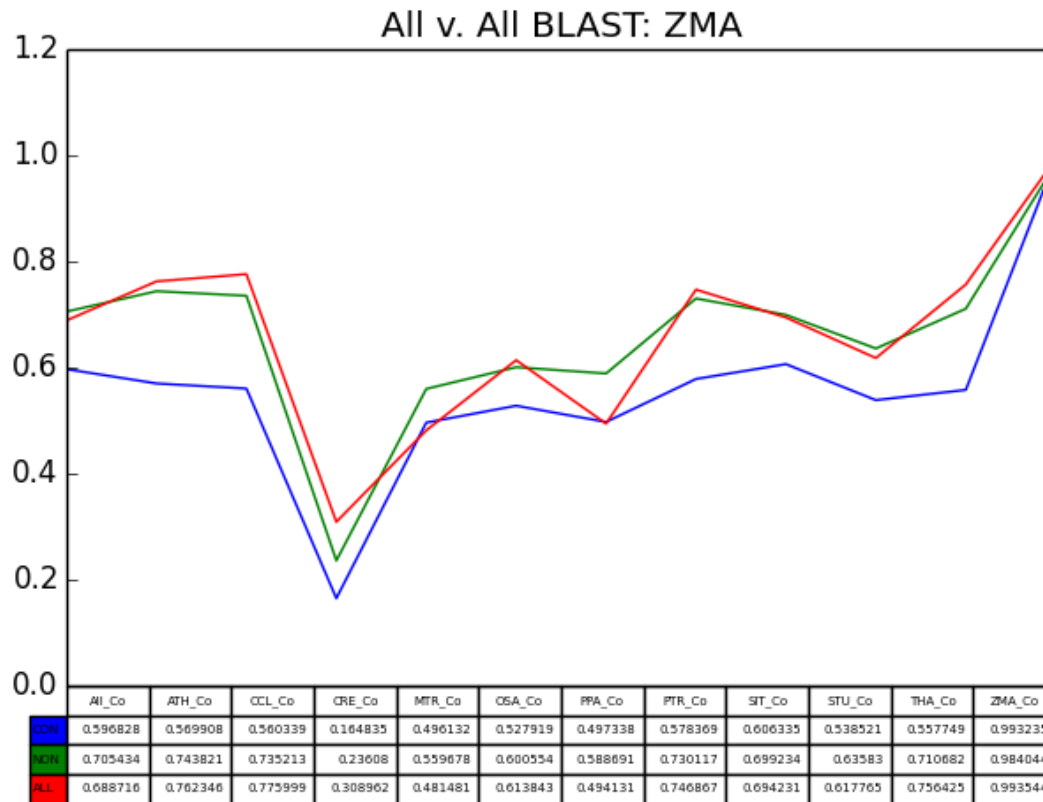
All vs. All BLAST



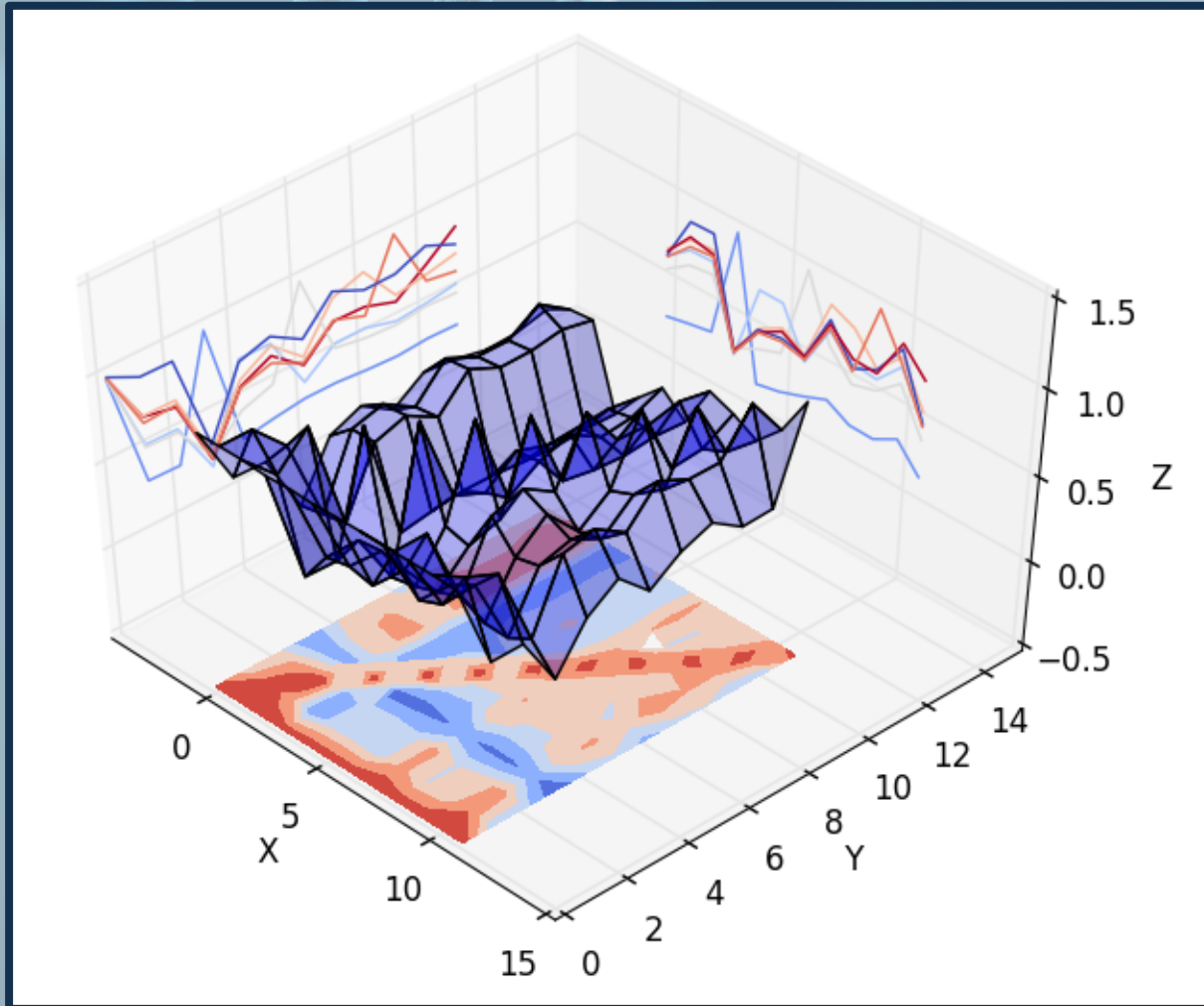
Results



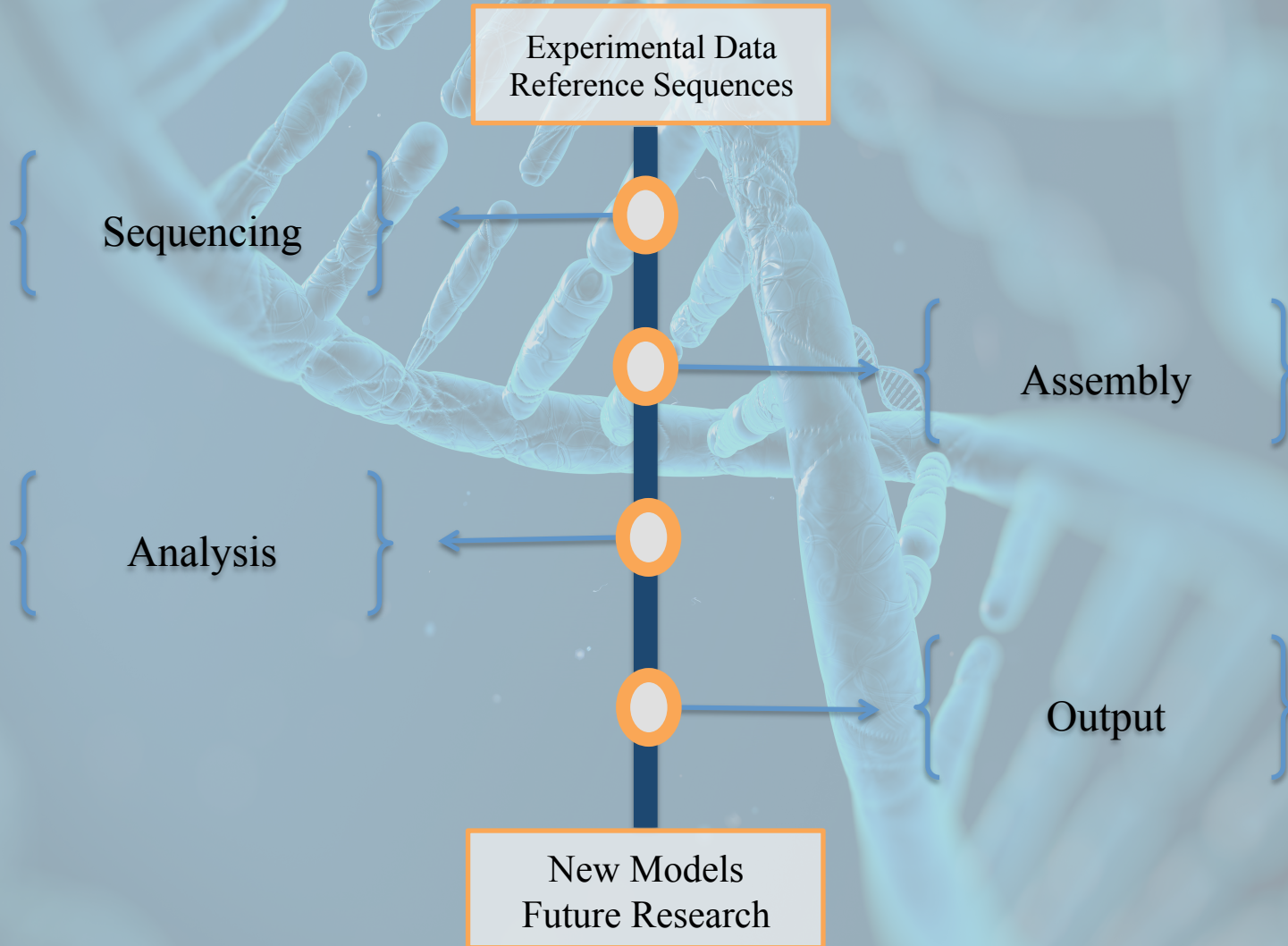
Results



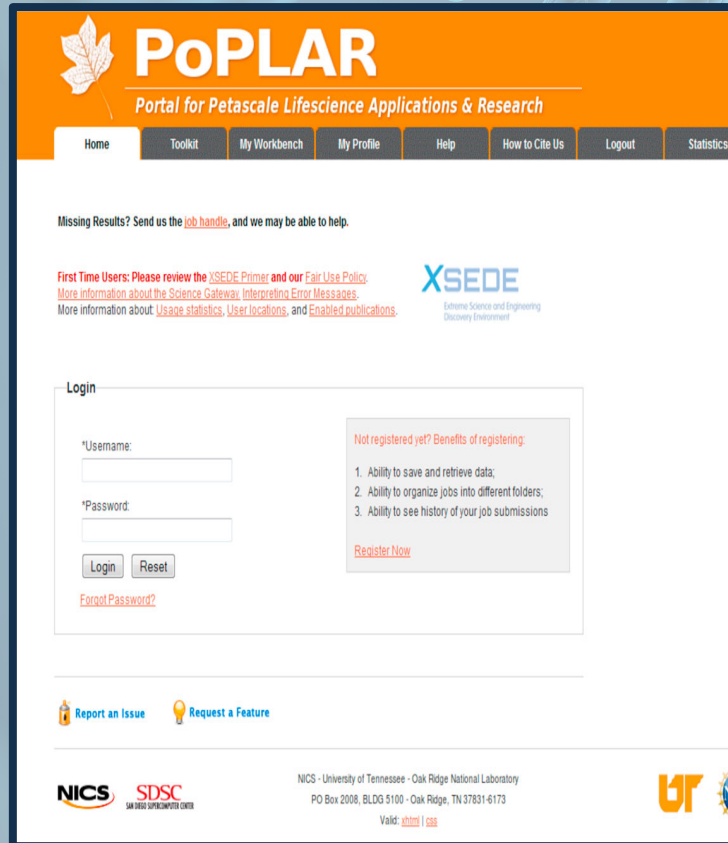
Results



Science Gateways



PoPLAR



The screenshot shows the PoPLAR home page. At the top, there is an orange header with the PoPLAR logo (a leaf) and the text "PoPLAR Portal for Petascale Lifescience Applications & Research". Below the header is a navigation menu with buttons for Home, Toolkit, My Workbench, My Profile, Help, How to Cite Us, Logout, and Statistics. The main content area includes a message about missing results, a section for first-time users with links to XSEDE resources, and a login form. The login form has fields for "Username:" and "Password:", "Login" and "Reset" buttons, and a "Forgot Password?" link. A registration box on the right lists benefits of registering and a "Register Now" link. At the bottom, there are links for "Report an Issue" and "Request a Feature", and logos for NICS, SDSC, and the University of Tennessee.



The screenshot shows the PoPLAR job configuration page. The header and navigation menu are identical to the home page. The "My Workbench" tab is active, showing a "Folders" section with a tree view containing "Folder1", "Data (2)", and "Tasks (86)". The main area is titled "Task Summary" and shows configuration for a job named "HSP BLAST on Kraken: Highly Scalable Parallel BLAST run on Kraken (B. Rekepalli, A. Vose, and P. Giblock)". There are sections for "Simple Parameters" and "Advanced Parameters". The "Simple Parameters" section includes fields for "Runtime (Hours)" (set to 10) and "Number of nodes (12 cores per node)" (set to 2000). The "Advanced Parameters" section includes fields for "BLAST output format (7, 8, or 9)" (set to 7), "Expectation value (E) threshold for saving hits" (set to 10), and "Which type of BLAST?" (radio buttons for blastp and blastn, with blastp selected). There are "Save Parameters", "Reset", and "Cancel" buttons at the bottom of the configuration area. At the bottom of the page, there are links for "Report an Issue" and "Request a Feature", and logos for NICS, SDSC, and the University of Tennessee.

Rekepalli et al.: PoPLAR: Portal for Petascale Lifescience Applications and Research

Future Goals

- Extension of HSP-BLAST all v. all analysis
- Improved automated workflows
- HSP implementation of additional tools
- Increased accessibility for computational tools through science gateways

References

Gerstein, M. B. The real cost of sequencing: higher than you think!. *Genome Biology*, 125.

Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 499-511.

Hunter, J. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 9, 90-95.

MACH 1.0 - Markov Chain Haplotyping. (n.d.). *MACH 1.0 - Markov Chain Haplotyping*. Retrieved August 4, 2014, from <http://www.sph.umich.edu/csg/abecasis/MaCH/>

Moreno-Hagelsieb, G., & Latimer, K. (2007, November 26). Choosing BLAST options for better detection of orthologs as reciprocal best hits. . Retrieved August 4, 2014, from

Rekapalli et al.: PoPLAR: Portal for Petascale Lifescience Applications and Research. *BMC Bioinformatics* 2013 14(Suppl 9):S3.

What is a gene?. (n.d.). *Genetics Home Reference*. Retrieved August 4, 2014, from <http://ghr.nlm.nih.gov/handbook/basics/gene>

Questions?

Acknowledgements:

