Newsletter

Spring 2018

Center for HIGH PERFORMANCE COMPUTING
THE UNIVERSITY OF UTAH

Research Highlight

Predicting Trait Values and Evolutionary Change from DNA Sequences

by Zachary Gompert, Utah State University Department of Biology

How predictable is evolution? This question has been asked and answered in various ways. Studies of parallel and convergent evolution have shown that species can predictably evolve similar phenotypes in response to similar environmental challenges, and that this sometimes even involves the same genes or mutations. On the other hand, scientists have argued that major external phenomena, such as cataclysmic meteor strikes and climate cycles, render long-term patterns of evolution unpredictable. Thus, evolution can be predictable to different degrees depending on the scale and specific features one is interested in.

The *Gompert lab at Utah State University* thinks a lot about predictability, both in terms of the predictability of evolution, and in terms of predicting phenotypes (i.e., trait values) from genetic/genomic data. In other words, we want to be able to predict traits from genes, and to predict how such traits and the underlying gene/allele frequencies change. And when we can't do these things, we want to understand why. Our work often relies on computationally intensive statistical modelling and simulations, which we use both to develop theory and to fit models. This requires access to large numbers of compute nodes, and in some cases large amounts of memory, substantial disk space and long running jobs, all of which have been made possible by USU's partnership with the University of Utah CHPC (UofU CHPC). Here I will outline some of our recent work that has been facilitated by the computational resources at the UofU CHPC.

Predicting traits (butterfly wing pattern) from DNA sequences

Predicting phenotypes (trait values) from genetic data is a key goal in biology; indeed, predicting traits from DNA sequences was one of the five grand challenges in biology recently articulated by NSF. Despite considerable efforts, it is still hard to predict trait values from genetic data, particularly for complex or quantitative traits. This difficulty arises from the fact that many genetic loci often contribute to trait variation and this frequently includes many rare genetic variants, genetic loci with small effects, or genetic loci with effects that depend on the environment or genetic background in which they are found. Understanding the genetics of complex trait variation within and between species is particularly difficult, as it necessitates genetic mapping in structured populations, which can confound attempts to identify causal genetic variants. Statistical models, such as recently developed Bayesian models for variable selection and genomic prediction (e.g., Bayesian sparse linear mixed models, Zhou *et al.* 2013) exist to overcome some of these issues. But such methods involve fitting hierarchical Bayesian models with hundreds of thousands to millions of model parameters using Markov chain Monte Carlo techniques, which necessitates substantial computational resources.

We recently used CHPC resources to apply such models and methods to generate genome-estimated trait values for many wing pattern characters for over a thousand *Lycaeides* butterflies sampled from multiple closely related species (see Figure 1). Using these genome-estimated trait values, we were then able to apply evolutionary quantitative genetic methods to quantify the role of genetic constraints in shaping patterns of wing pattern evolution.

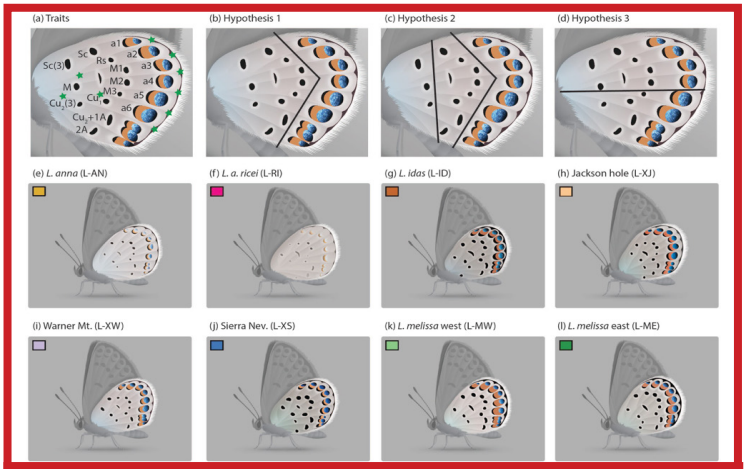

Figure 1. Stylized drawings of wings.

(a) shows the pattern elements measured. Size and position were measured for the labeled elements, and stars denote additional landmark positions along wing veins. **(b)--(d)** show our modularity hypotheses where black lines divide the wings into sections; pattern elements within a section are constrained by a shared genetic basis while elements across sections are not. **(e)--(l)** show drawings from each taxon that are meant to highlight key differences in wing patterns among groups. *Drawings by Amy Springer.*

We generated genome-estimated trait values for 69 wing pattern characters (the sizes and positions of wing pattern spots and veins) for >1000 butterflies, with analyses repeated at different taxonomic scales (there were 828 mapping analyses total). This was done with the free, open source program gemma, which implements computational approaches to fit Bayesian sparse linear mixed models. Each analysis generated predictions based on approximately 60,000 genetic markers (single nucleotide polymorphisms or SNPs) and required 5 million Markov chain Monte Carlo steps (five chains with 1 million iterations each). This approach and algorithm provide a means to numerically integrate and sample from the posterior probability distribution of these model parameters, which in this case, is a space with >100,000 dimensions. Running these analyses, which were spread across dozens of compute nodes and cores, required approximately 10,000 computer hours or over 400 computer days. Such analyses would be intractable without access to a large-scale computer cluster like the UofU CHPC. By combining the output from these runs, we were able to generate genetic variance-covariance matrices for the set of wing pattern traits, which we could then subject to thousands of random selection vectors to assess the extent to which the evolution of these traits was constrained by genetic covariances, and whether such within species constraints predicted patterns of among species divergence.

We found that wing pattern was polygenic with mostly minor effect loci. We identified conserved modules of integrated wing pattern elements within populations and species, and showed that trait covariances within modules have a genetic basis, and thus represent genetic constraints that can channel evolution (Figure 2). Consistent with this, we found evidence that evolutionary changes in wing patterns among populations and species occurred in the directions of genetic covariances within these groups. Thus, we were able to show that genetic constraints affect patterns of biological diversity (wing pattern) in *Lycaeides*, and provide an analytical template for similar work in other systems. Our paper describing these results is coming out soon in a special issue on trait mapping in Molecular Ecology Resources (Lucas *et al.*, 2018).

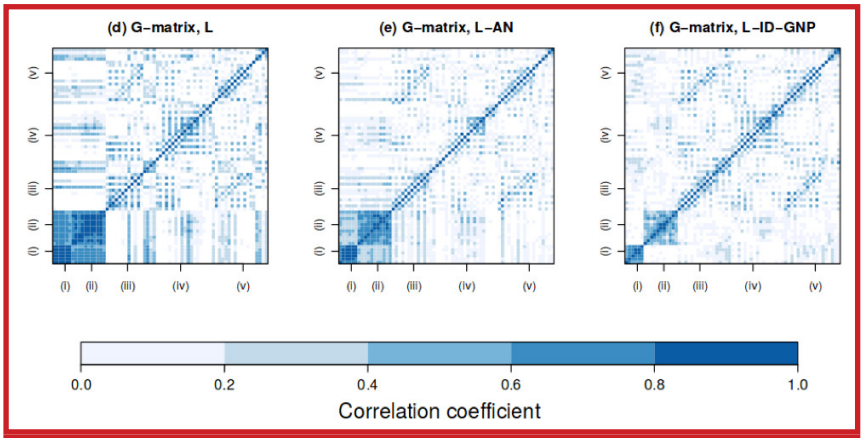

Figure 2. Heat map of example genetic correlation matrices.

L = *Lycaeides* species complex, **L-AN** = *Lycaeides anna*, and **L-ID-GNP** = *L. idas,* Garnet Peak population. Roman numerals denote different sets of traits: **(i)** = orange spot size, **(ii)** = black spot size, **(iii)** = orange spot position, **(iv)** = spot position, and **(v)** = wing vein position.

Natural selection and predicting evolutionary change

One expects evolutionary change to be more predictable when it is driven mostly by natural selection. Thus, one key component of our research is trying to estimate and parse the contribution of selection to evolutionary change, particularly from evolutionary time series data. Evolutionary time series data include trait of gene/allele frequency data from multi-generational lab or field experiments or natural populations sampled across populations, as well as data associating genotypes or trait values with survival or other fitness components within generations (i.e., from mark-release-recapture experiments). We have applied such methods to stick insects, seed beetles, butterflies and simulated data. To infer selection from time series data we frequently use approximate Bayesian computation.

In standard Bayesian inference, the likelihood of the data given some model parameters is combined with prior probabilities of the parameters to yield a posterior distribution, that is a multi-dimensional probability distribution specifying the probabilities of different parameter values for all of the model parameters conditional on the data. For some models, the likelihood function is either unknown or not easily computed. In such cases, approximate Bayesian computation can be used, as it replaces the likelihood function with an evaluation of data simulated under the model. Specifically, parameter values for the model are drawn from their prior distributions, these are then used to parameterize a simulation of evolution, and finally summaries of the evolutionary time series from the simulations are computed and compared to the real data. Parameter values from the simulations that best recreate some aspects of the true data are then used to learn about or inform our knowledge of the true posterior distribution. This approach allows us to tailor simulations to the particular details and history of an evolution experiment, including details of the data-generation process.

Approximate Bayesian computation requires many simulations to generate even a few combinations of model parameter values that generate output similar to the real data. For example, an ongoing project with seed beetles in our lab has relied on approximately 1 billion simulations of evolution for a 16 generation time series. Even when each simulation is relatively quick, this requires substantial computational time and effort. On the other hand, these simulations are easy to parallelize and thus have benefited greatly from the large numbers of nodes and cores on the CHPC.

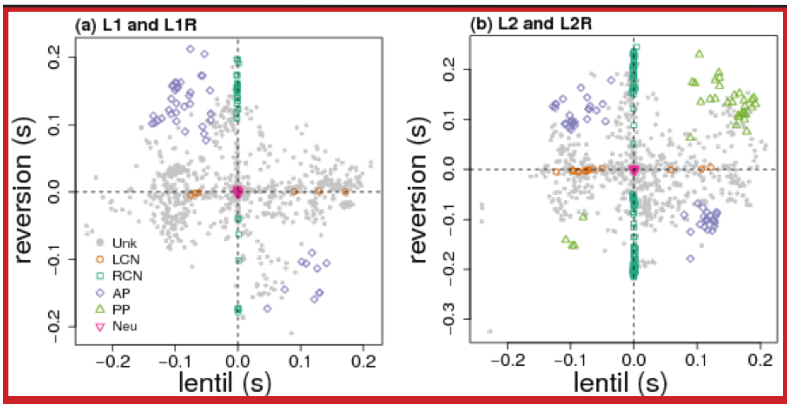

Figure 3. Estimates of selection on individual genes in a novel (lentil) and ancestral (reversion) host.

Points are colored to reflect whether the evolutionary dynamics for each gene most likely reflect selection in lentil only (LCN), selection in reversion only (RCN), opposing selection between hosts (AP), similar selection on both hosts (PP), neutral evolution (Neu), or unknown or uncertain processes (Unk).

Using approximate Bayesian computation we have shown that adaptation to a novel host in seed beetles involves genetic trade-offs, such that genetic variants favored on the novel host (in this case lentil beans) were selected against on the ancestral host (in this case mung beans) (Figure 3; Gompert & Messina, 2016). Such trade-offs enhance the predictability of evolution. And in a paper our group published in a recent issue of Science, we used approximate Bayesian computation to show how selection in stick insects varies across different life history stages (Nosil et al., 2018). This variation increases the complexity and thereby reduces the predictability of evolution. These simulation-based inference of the evolutionary process allow one to consider models of arbitrary complexity and to let the process of interest dictate the model rather than trying to cram data into a simpler analytical framework. By using these approaches, we have begun to identify key determinants of the predictability of evolution, and have done so in a way that can be scaled up to larger and larger genomic data sets.

References

Gompert, Z., Messina, F. J. (2016). Genomic evidence that resource-based trade-offs limit host-range expansion in a seed beetle. *Evolution*, 70(6), 1249-1264.

Lucas, L., Nice, C., Gompert, Z. (2018) Genetic constraints on wing pattern variation in *Lycaeides* butterflies: a case study on mapping complex, multifaceted traits in structured populations. *Molecular Ecology Resources*, doi pending.

Nosil, P., Villoutreix, R., de Carvalho, C., Farkas, T., Soria-Carrasco, V., Feder, J., Crespi, B., Gompert, Z. (2018) Natural selection and the predictability of evolution in *Timema* stick insects. *Science*, 359, 765-770

Zhou, X., Carbonetto, P., Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS genetics*, 9(2), e1003264.

CHPC Summer 2018 Presentation Schedule

All presentations are held in the INSCC Auditorium, starting at 1pm.

*** = 1 hour**

**** = 2 hours, Hands -on**

***** = 9am-3pm, w/ break for lunch**

See *https://www.chpc.utah.edu/presentations/* for details on trainings.

Date	Presentation Title	Presenters
May 17	Introduction to HPC*	Anita Orendt
May 22	Introduction to Linux, Part 1**	Brett Milash & Wim Cardoen
May 24	Introduction to Linux, Part 2**	Brett Milash & Wim Cardoen
May 31	Introduction to Linux, Part 3**	Brett Milash & Wim Cardoen
June 4-7	XSEDE Summer Bootcamp ***	Wim Cardoen
June 12	Introduction to Linux, part 4**	Wim Cardoen & Brett Milash
June 14	Module Basics*	Anita Orendt
June 19	Slurm Basics*	Anita Orendt
June 21	Introduction to Python, Part 1**	Brett Milash & Wim Cardoen
June 26	Introduction to Python, Part 2**	Brett Milash & Wim Cardoen
June 28	Numpy/Scipy**	Brett Milash & Wim Cardoen
July 10	Using Git*	Robben Migacz

CHPC has developed a series of courses to help users make the most of CHPC resources. During spring and summer semesters we present an abbreviated set. There is no cost associated with these training sessions. There is no need to register, with the exception of the XSEDE Workshops. Also, CHPC will be moving to using Zoom for remote attendance to the presentations.

Facilitation Highlight

Archaeological Sites as Endangered Species: Using Next Generation Models to Predict and Protect Cultural Properties on the Grand Staircase-Escalante National Monument

by Peter M. Yaworsky, Brian F. Codding, Kenneth B. Vernon, and Wim R. Cardoen, The University of Utah

The Objective

As part of a broader Class I project for the Bureau of Land Management (BLM), the University of Utah Archaeological Center (UUAC) was contracted to create a statistical model for predicting the likelihood of archaeological sites, also referred to as cultural resources, across the Grand Staircase-Escalante National Monument (GSENM). This Cultural Resources Predictive Model (GSENM-CRPM) uses a complete sample of all known archaeological sites broken into time period specific components to predict unknown sites based on environmental characteristics associated with these sites using a species distribution model (following a Maximum Entropy, MaxEnt, approach). The result of this is a set of robust statistical models capable of predicting the occurrence of cultural resources throughout the region.

The UUAC director, Dr. Brian Codding and University of Utah Anthropology graduate students Peter Yaworsky and Kenneth B. Vernon began developing a MaxEnt model specific to archaeological sites in August 2017. When looking to generate these models, the group realized they did not have the computational resources to accomplish the task and they reached out the Center for High Performance Computing (CHPC), with Peter acting as the primary contact. During an initial meeting with Dr. Anita Orendt, CHPC's Research Consulting & Faculty Engagement Coordinator and ACI-REF, Peter discussed the computational needs of the project. The memory and CPU of the statistical calculations (using the R statistical programming environment) were beyond the available resources on their office computers, and even if they were able to perform the simulations on these resources it would take too long and they would not meet the project deadline. However, there were barriers to moving this research to CHPC as Peter had no experience using Linux and HPC.

The Solution

At the end of Peter's initial meeting, Anita introduced Peter to CHPC Scientific Consultant and ACI-REF Dr. Wim Cardoen, who, in his current role, takes care of the R statistical package at CHPC in the broad sense -- he teaches an introductory class on R, performs the installation of R (core and external packages) on the CHPC clusters, writes the corresponding SLURM scripts, and consults with the CHPC user base when they have R-related questions. Peter then met with Wim to gain some familiarity with the CHPC clusters, specifically working in a Linux environment and using a batch system to submit his analyses to the cluster versus running them directly on his Windows computer.

In a first step, Wim installed eight external R packages, required to process spatial data. Among these packages were *rgdal* for spatial data processing and *rgeos* for vector processing, which are R interfaces to C libraries of gdal (geospatial abstraction layer) and geos (Geometry Engine - Open Source), respectively. Wim first installed the C libraries, and then had to address the fact that they were installed in non-default locations. Along with the R packages, Wim also installed the MaxEnt package on the clusters.

In addition, Peter shared his R code and data sets with Wim. Wim modified part of Peter's original code to make it more suited to run on CHPC's HPC clusters. He also tinkered with Peter's R code to determine the optimal use of the compute nodes. Wim found that the use of multiple cores (using the environment variable OMP_NUM_THREADS) did not significantly improve the performance of the code. Therefore, he decided to proceed by running multiple serial simulations per node, in order to maximize the efficient use of multi-core compute nodes. However, when using one simulation per core, Wim realized that the memory needs of each simulation placed an additional constraint on the runs, as each simulation required about 6 GB, a quantity greater than the memory per core of most of the CHPC nodes. Therefore the decision of the number of simulation per node was determined by the memory of the node. With this knowledge, Wim created the corresponding Slurm scripts for Peter, allowing him to proceed with the validation of the data, the generation of the models, and the use of CHPC resources to create the predictive rasters, as described below.

The original data set consisted of 132 geospatial predictor variables, called rasters. The geospatial rasters fell into five categories: resource distribution, climate, environmental productivity, landscape and soil attributes. Only 110 of the 132 geospatial rasters were used due to the abundance of missing values for 22 rasters.

The initial calculations tested whether the sample areas (areas inventoried for archeological sites) were adequately represented by parameters derived from the 110 predictor rasters. Wim assisted Peter in running these preliminary calculations, which were finished on CHPC resources within several days.

In the subsequent calculations the MaxEnt method was used to produce and analyze two types/generations of models. The MaxEnt method allowed for the determination of the relative weight of the different predictor rasters. In the first generation,

37 models were created using all 110 predictor rasters. Each of these 37 models had a certain time frame associated with it. The second generation of models was a further refinement by selecting the most important predictor raster variables and by dropping the predictor variables with a strong correlation to the selected predictors. In the final stage the five refined models were used to create the four predictive raster for different time periods, namely for Archaic, Formative, Late Prehistoric, and Historic Time Periods; these were then combined, taking the average of these four individual time periods to produce the General Time Period predictive raster and taking highest probability from each of these time periods, resulting in the Combined Time Period predictive raster. (see Figure 1).

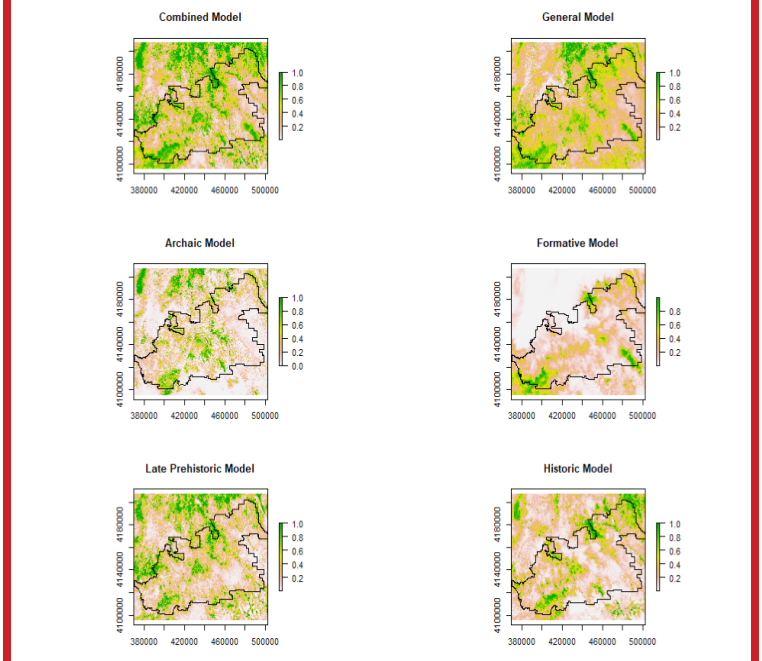

Figure 1. Final predictive rasters of the region studied.

The black outline represents the boundaries of the GSENM. The x- and y-axes are in meters, using the NAD83 UTMs zone 12 coordinates. The color scale represents a probability finding an archaeological site in a given cell.

The Result

The first training run of MaxEnt resulted in 37 predictive models based on the 110 predictor rasters and all 4400 archaeological sites. These models allow one to predict where, for example, a residential site dating to the Archaic period, or a rock art site from the Late Prehistoric period is likely to be found within the Monument. The second training run of MaxEnt resulted in four new time period specific predictive models and one general time period predictive model. These models differ from the preliminary time period models in that they utilize a subset of the 110 raster variables that did not covary with one another.

The refined models were used to create six predictive rasters or maps showing the probability of site occurrence (from 0 to 1) at a 5 m^2 resolution. They include the four specific time period rasters (Archaic, Formative, Late Prehistoric, and Historic), one General time period raster, and one Combined time period raster. The Combined raster was created by overlaying the four time period rasters and keeping the highest cell values. Where the General time period raster identifies *only* locations where sites affiliated with specific time periods are likely to occur *together*, the Combined time period raster identifies *any* location where sites affiliated with specific time periods are likely to occur *together or separately*. Thus, the UUAC project was capable of addressing the longstanding problem of underestimating the potential for archaeological resources that accompanied the more promiscuous lumping strategies of previous modeling efforts.

The research, therefore, has both intellectual merit and broader impacts. First, it furnishes anthropology and archaeology with a new method for evaluating hypotheses regarding the evolution of human land-use through time. Second, the project provides a stepping stone to future research aimed at addressing questions of prehistoric land-use on a regional scale. Finally, it equips federal land managers with a powerful new tool, allowing them to craft more effective preservation strategies on public lands.

Changes to the Allocation Process

Starting April 1, the general allocation awards are for time on kingspeak and notchpeak instead of on ember and kingspeak. Ember general nodes will join lonepeak and tangent as unallocated resources. While this change will result in a net lowering of the number of core hours available for allocation, the improvement in performance of the notchpeak nodes versus the ember nodes, will result in a gain in computational power. The maximum award has been decreased from 250,000 core hours to 200,000 for regular, and from 30,000 to 20,000 for quick allocations.

The second change is in the allocation process itself. Starting with allocation requests made during the current quarter, the quick allocation form will be simplified. In addition, regular allocation requests for 20,000 core hours or less will also use the simplified form. The simplified form will require only the following information: PI, title, abstract, sources of funding, publication based on CHPC usage. This new form will be available for use before the request for Summer 2018 allocations is made.

In addition, the protected environment HPC resource (Redwood PE), which is currently running in an unallocated manner, will be moving to an allocation process starting in July. In this case, the allocation process is slightly different, in that priority will be given to NIH funded projects.

Technical Highlight — New Protected Environment

by Anita Orendt

As some of you already know, CHPC is nearing completion of a refresh of the protected environment (PE). This refresh was made possible due to the award of a NIH Shared Instrumentation Grant, Grant number 1S10OD021644-01A1, in April 2017. The award allowed CHPC to deploy a complete refresh of the existing PE, and in the process expand the capabilities and increase the security relative to the initial CHPC PE deployment. In addition, the refreshed PE is configured to allow for expansion in a condominium fashion, in both the storage and in the HPC components. The different components of the new PE were made accessible to users as they were deployed, most during the first quarter of 2018.

The refresh of the PE includes:

- *HPC Cluster*
- *Home Directory*
- *Project Space*
- *New firewall*
- *Archive Storage*
- *Scratch Directory*
- *Windows Server*
- *VM farm*
- *New Security Information & Event Management (SIEM) solution*

As required by the award, CHPC has also formed a PE Policy & Allocation Committee to oversee the deployment and the subsequent use of this resource. Members include Orly Alter, Bioengineering/SCI; Chris Butson, Bioengineering/Neurology/SCI; Thomas Cheatham, PI (ex-officio); Julio Facelli, BMI; Cynthia Furse, ECE/VPR; Bryan Jones, Moran; Bernie LaSalle, BMI; Tim Parnell, HCI; Anita Orendt, CHPC (Chair); Aaron Quinlan, Human Genetics.

With the new PE, there are changes in the policies and processes for usage of the PE. Below, both a description of the new resource and the new policies are given.

HPC:

The new PE HPC cluster is called Redwood. As with the clusters in the general environment, the power and networking infrastructure is in place to expand the cluster in a condominium manner.

The initial, general cluster hardware includes 17 compute nodes:

- *4 Intel XeonSP (Skylake) nodes each with 32 cores and 192 GB RAM (128 total cores)*
- *11 Intel Broadwell nodes each with 28 cores and 128 GB RAM (308 total cores)*
- *2 GPU nodes with 4 x GTX1080Ti GPUs and 32 cores each (Intel XeonSP), 192 GB RAM*

In addition, there are two general interactive (login) nodes (XeonSP, 32 core), and an EDR infiniband interconnect fabric.

Whereas, the HPC usage was run unallocated in the old PE, in the new PE there will be an allocation process for time on the general compute nodes with priority given to projects with NIH funding. The PE Policy & Allocation committee will be the group reviewing applications and making award recommendation. General nodes left idle will be available for use in the freecycle mode, with preemption.

Along with the general nodes there is the ability to add owner nodes, both as compute and interactive nodes. There have already been 56 owner compute nodes added to redwood. Owner nodes left idle can be utilized by all PE users in the guest mode, again with preemption.

Storage:

The storage that will house the home directories, project spaces, and scratch is named Mammoth. The scratch space, /scratch/mammoth/serial, has a capacity of 160 TB, while the total initial capacity for home and project spaces is another 160 TB.

All users get a 50 GB home directory (*/uufs/chpc.utah.edu/common/PE/UNID*). There will be no increases in the quota for this space. This is backed up on a nightly incremental, weekly full schedule with a 2 week retention window. This space should be used for user specific files such as scripts. The 50 GB is a soft quota, with 75 GB being the hard limit. When a user exceeds 50 GB of usage, as long as they do not exceed 75 GB, a 7-day clock is started. If the user cleans up their usage to below 50 GB before the end of this 7-day window, their usage will not be interrupted. Any usage that exceeds 75 GB, or staying above 50 GB for longer than 7 days, will result in the user no longer being able to write to their home directory.

Each project will be provisioned, by default, with a 250 GB project space (*/uufs/chpc.utah.edu/common/PE/project*). Access to this space is limited to the users working on the project. If the project has an IRB, the users must be listed on the IRB. If the project is not governed by an IRB, then the PI of the project will need to approve a user before CHPC will provide access. For projects that need more than 250 GB: if the project is NIH funded, the PI can make a request for up to 5 TB, with justification of need. For non-NIH funded projects, or for NIH projects needing more than 5 TB, additional storage can be purchased at a cost of $150/TB. This space will be grown as needed. While this space will initially be backed up, as was the project space in the old PE, as this space grows and as CHPC develops a new backup strategy, this will change.

Along with mammoth, there is also an archive storage, elm, with an initial capacity of 1 PB; NIH funded projects get 1 TB/project free. Space on this file system is available for $120/TB. Again, this space will be grown as additional capacity

is needed. CHPC anticipates that this storage will become the place for backup of project space data.

VM farm:

The new VM farm is named Prismatic. With the replacement of the VM hardware in the PE, VMs will no longer be free, unless the project that the VM is supporting is NIH funded. The VM pricing model is based on a block sizing increment, with five different block sizes available, with the cost per block based on the cost of the hardware and the number of blocks available to sell. The prices are for the warranty lifetime of the VM hardware, which was purchased with a 5-year warranty. CHPC will offer to deploy a VM for a trial period of up to 6 months free of charge, provided that the VM does not require substantial customization. Note that these prices are for internal, research needs; if the VM is for an external project the cost is based on the total cost of operation, which substantially raises the cost.

VM Description				
Blocks	***RAM (GB)***	***Cores***	***Storage (GB)***	***Price***
1	4	2	50	$350
2	8	2	100	$700
4	16	4	200	$1,400
8	32	8	400	$2,800
16	64	8	800	$5,600

Another change is that there are two different storage offerings on the VM farm: SSD storage which is not encrypted and self-encrypting 7200 RPM spinning drives. Unless encryption is needed, the SSD storage will be used. Additional VM storage is available, in 100GB increments, at a cost of $1000/TB for SSD storage and $300/TB for encrypted spinning storage.

Windows server:

The replacement for Swasey will be called Narwhal. While Narwhal is not yet available for use, in the new PE there will be changes made on access to the windows compute environment. Currently, Swasey is a single physical server being used for general access to the PE, for typical desktop applications (e.g., Word) and services, as well as for computational needs.

In the new setup, the different usages will be segregated onto different servers, both physical and virtual. There will be a set of gateway servers ("session host boxes") that provide users with access to the PE using remote desktop with DUO two factor authentication. Having multiple servers will allow for higher availability as they can be removed from service to be updated independently. From these session host servers, users with compute intensive needs will then connect to the new Narwhal compute server. This server has 24 physical cores, 512GB of memory, and 1TB of local SSD space, and will have installations of the statistical packages currently found on Swasey. It also allows for mounting of the PE home and project spaces if needed. By isolating the compute functionality to a server without direct login access, the need for applying OS updates immediately is mitigated, allowing for longer run times between updates.

As the configuration of the windows environment is completed more information will be shared via the CHPC website and announcements to the PE user mailing list.

RMACC HPC Symposium 2018

The 2018 Rocky Mountain Advanced Computing Consortium (RMACC) High Performance Computing (HPC) symposium will be held on the University of Colorado Boulder campus August 7-9. This symposium features a wide array of panel discussions, technical presentations, and tutorial sessions on research, education, and best practices in the areas of computational science and high performance computing.

As part of the conference, there is a student track and a student poster competition, with the presenters of the winning posters being awarded travel to SC18 which will be held Dallas, November 11-16.

Watch for registration and poster submission in May!

Human Genomics, dbGaP, and the Use of the Protected Environment

In March 2015, the NIH published security best practices for controlled access data that was subject to the NIH Genomic Data Sharing (GDS) policy. This included data in the NIH database of Genotypes and Phenotypes (dbGaP). In response to this, CHPC published a set of guidelines: *https://www.chpc.utah.edu/documentation/policies/1.6SecurityPolicy.php#Pol1.6.5*

The CHPC general environment does not meet the security best practices for dbGaP without the addition of the use of two factor authentication for users working with the restricted data and extended access control lists (ACLs) for restricting access to the data. Note also that the encryption requirements render the Ceph archive storage in the general environment unusable for any restricted data.

CHPC requests that all groups determine if their projects are governed by this NIH policy, and if so contact CHPC via *helpdesk@chpc.utah.edu* to transition to the use of the PE where these security practices are met by default.

In addition, as with the refresh of the PE, we now have the capacity to house all projects dealing with human genomic data in this restricted environment, we strongly encourage all projects that are working with human genomic data start to transition into the use of the PE; again you can contact CHPC via *helpdesk@chpc.utah.edu* to discuss these recommendations.

Center for
HIGH PERFORMANCE COMPUTING
THE UNIVERSITY OF UTAH

The University of Utah
University Information Technology
Center for High Performance Computing
155 South 1452 East, Room 405
SALT LAKE CITY, UT 84112-0190

Thank You for Using CHPC Resources!

Welcome to CHPC News!

If you would like to be added to our mailing list, please provide the following information and send via the CHPC contact methods listed below:

Name:
Phone:
Email:

Department or Affiliation:

Address: (U Campus or U.S. Mail)

Please help us continue to provide you with access to cutting edge equipment.

ACKNOWLEDGEMENTS

If you use CHPC computer time or staff resources, we request that you acknowledge this in technical reports, publications, and dissertations.

Example of what we ask you to include in your acknowledgements:

"A grant of computer time from the Center for High Performance Computing is gratefully acknowledged."

If you make use of the CHPC protected environment, please also acknowledge the NIH shared instrumentation grant:

"The computational resources used were partially funded by the NIH Shared Instrumentation Grant 1S10OD021644-01A1."

Please submit copies or citations of dissertations, reports, pre-prints, and reprints in which the CHPC is acknowledged in one of the following ways:

Electronic responses

Email: ***helpdesk@chpc.utah.edu***
colette.durrant@utah.edu

Fax: (801)-585-5366

Paper responses

U.S. Mail: 155 South 1452 East, Rm 405
Salt Lake City, UT 84112-0190

U Campus Mail: INSCC 405