Bioinformatic Analysis of Effector Protein Genes in *Bremia lactucae*

Yuji Mori, Kelsey Wood
Michelmore Lab
UC Davis Genome Center

Introduction

The **downy mildews** are plant pathogens that threaten many crops in commercial agriculture. These pathogens secrete **effector proteins**, which can suppress immune behavior in the host and render them susceptible. Selective forces act on the plant's defense mechanisms to resist the parasites, but in response, downy mildew effectors also evolve to avoid host detection and remain infectious. The result of this interaction is rapid diversification within both the effectors and plant immune systems, which creates challenges in their overall study.

Objectives

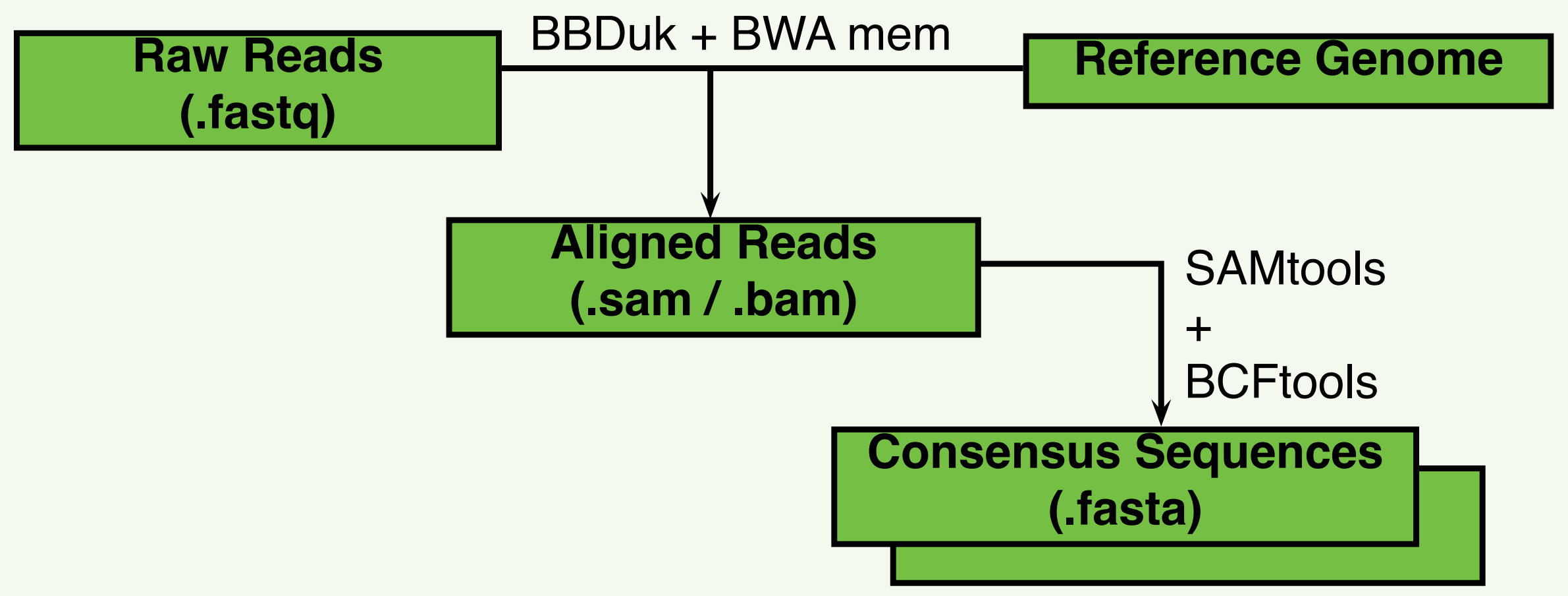
- Determine the evolutionary relationship between a set of 56 candidate effectors (WY genes).
- Quantify the selective forces acting on each region and interpret their implications on gene function.
- Correlate the presence/absence of genetic markers to observable patterns of infection on lettuce samples.

Preliminary Data Processing

Raw sequencing data for 95 isolates of *Bremia lactucae* were generated by the Michelmore lab. Sequences were trimmed and aligned to the 56 candidate WY regions against a de-novo assembled reference genome. The result was a collection of 95 BAM files, which were subsequently merged and sorted.

In order to convert these We employed a variant calling procedure to generate pairs of nucleotide consensus sequences using SAMtools and BCFtools.

The Consensus Calling Pipeline.



Raw Reads (.fastq)

BBDuk + BWA mem

Reference Genome

Aligned Reads (.sam / .bam)

SAMtools + BCFtools

Consensus Sequences (.fasta)

The sequences were further screened for premature stop codons and truncated appropriately. The processed data consists of 95 sets of 56 x 2 consensus sequences.

Bremia is a diploid organism, so we expect two sequences for each gene that are not necessarily identical. Currently, the consensus results are *un-phased*, meaning the variants are arbitrarily distributed between each allele.

Diversification Analysis

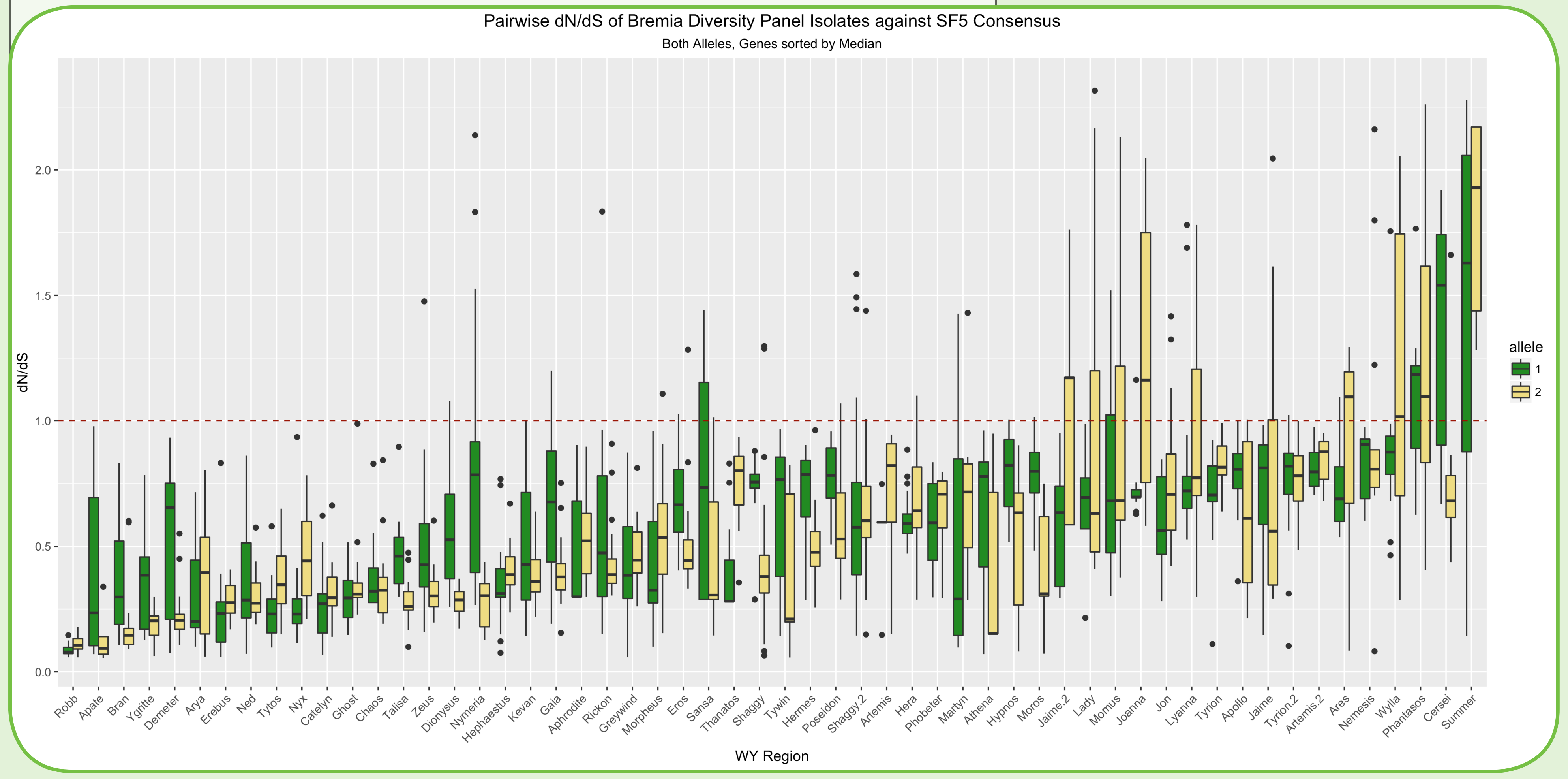
For each isolate, we performed multiple sequence alignment using the CLUSTALW algorithm and generated phylogenies to visualize the possible relationships between WY genes. Some candidate effector genes were suspected to be paralogous, so we used the Geneious software to more closely inspect any SNP variaton.

The direction and magnitiude of selection acting upon each protein-coding WY gene was estimated using the **dN/dS** (or Ka/Ks) statistic:

$$\frac{\text{non-synonymous substitutions per non-synonymous site}}{\text{synonymous substitutions per synonymous site}}$$

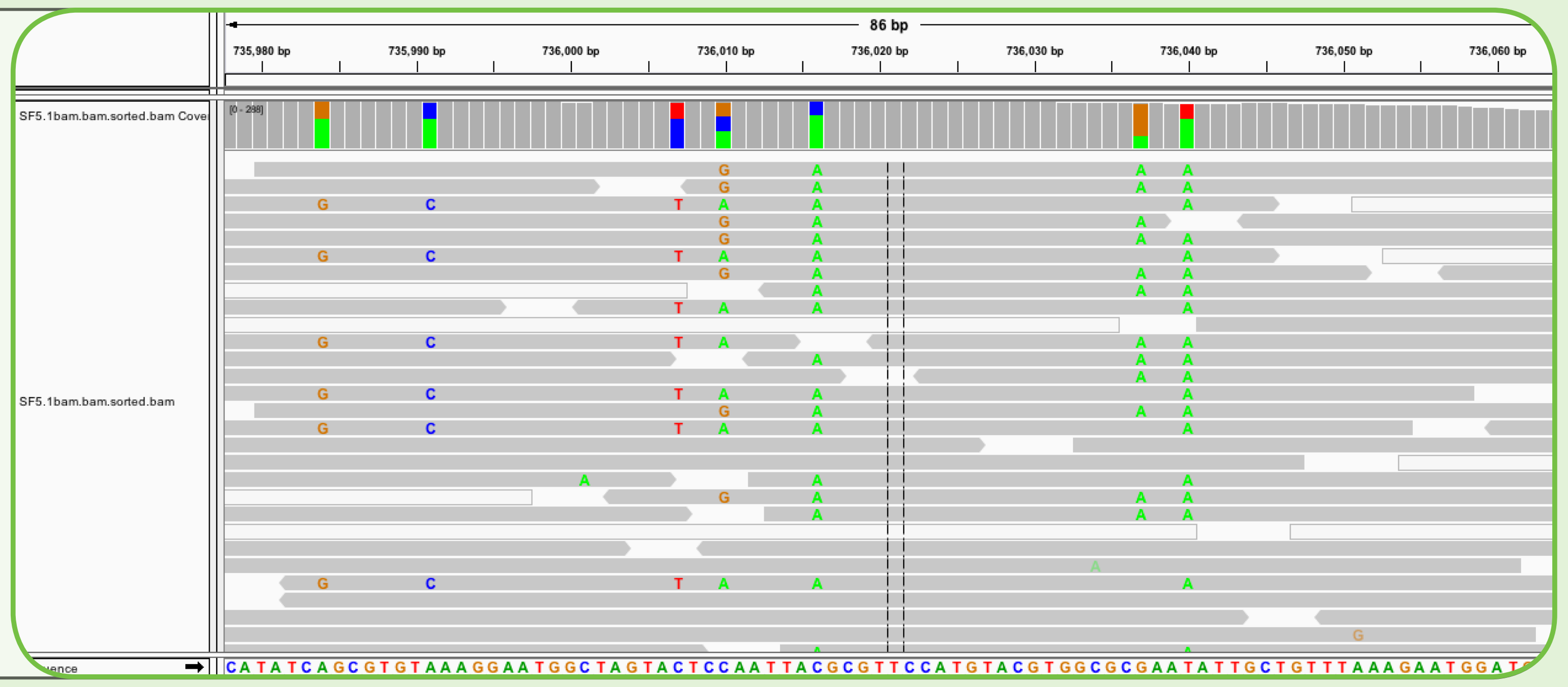
Sequences were pairwise codon-aligned to a baseline sequence from the SF5 isolate, then dN/dS calculations were made in PAML. Most genes are expected to be conserved (purifying selection) with a ratio below 1, but analysis showed that some regions undergo positive selection of new mutations.

Using regional data for each isolate, we later explored the differences in selective forces between isolates found in Europe compared to those from the United States.



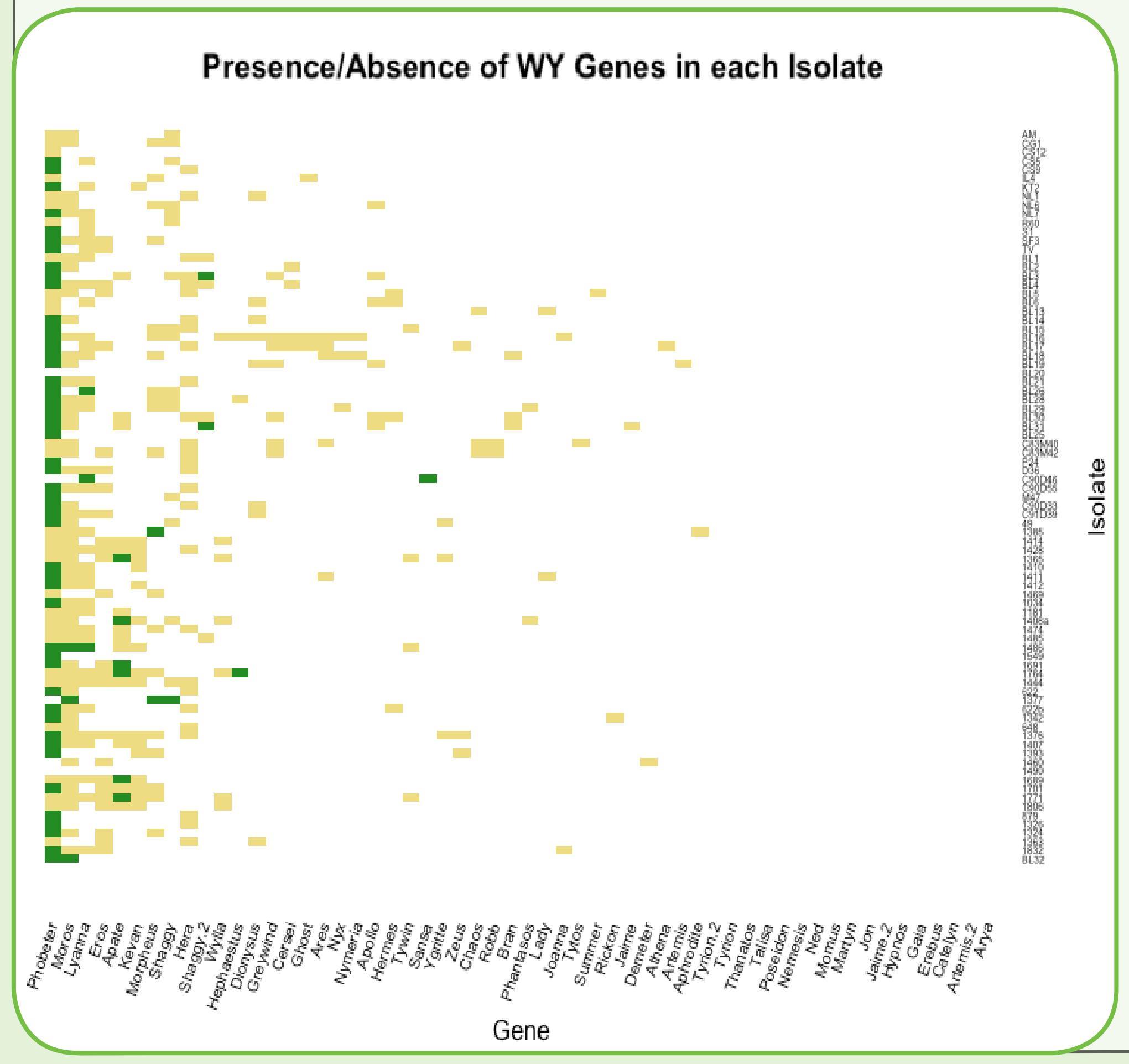
Visualization of a High-Coverage BAM File in IGV.

SF5 Isolate aligned to SF5 Reference Genome. Variants are automatically detected and colored.



Association with Phenotype Data

Consensus nucleotide sequences with premature stop signals or frameshift mutations typically produce a non-functioning protein. As such, each isolate can be represented as a set of "present", "partially present", or "absent" effector genes. Most genes are conserved across all isolates, but we found some cases with high evidence of degeneration.



Cramer's V Association between Gene Presence/Absence and 5 Phenotypes

	DM1 Lednicky	DM2 UCDM2	DM3 Dandie	DM4 R4T57D	DM5/8 Valmaine
Apate	0.111336215	0.049053448	0.10165559	0.05259254	0.202444083
Apollo	0.004756515	0.037152142	0.251221975	0.188982237	0.1
Ares	0.027605909	0.059681636	0.041292279	0.161780507	0.072547625
Dionysus	0.104922187	0.11542609	0.089026774	0.205498734	0.108739709
Eros	0.14472239	0.007021095	0.047476491	0.305714286	0.10394023
Greywind	0.224622148	0.285267338	0.009756359	0.079850937	0.170876686
Hera	0.059590997	0.063515947	0.182031403	0.170327592	0.093494699
Kevan	0.142753727	0.101261973	0.042074427	0.246317415	0.182217247
Lyanna	0.223371793	0.213714867	0.159102701	0.011293849	0.115727512
Moros	0.209158825	0.064187819	0.12225138	0.166627959	0.098978821
Morpheus	0.113573344	0.06973422	0.081770411	0.106313835	0.095971487
Phobeter	0.125725111	0.35975903	0.279391575	0.169682326	0.322748612
Shaggy	0.059456437	0.103312223	0.130926531	0.279108914	0.044194174
Wylla	0.014777773	0.018138386	0.108539491	0.01526562	0.108739709

Genes with sufficiently distributed data in each of the categories (present/partial/absent) were tested for association against 19 binary phenotype measurements. The statistic used is Cramer's V, where possible values range from 0 (no association) to 1 (complete association).

Discussion and Future Efforts

Multiple sequence alignment and tree building revealed consistent relationships and evolutionary distances between WY regions. However, close-up analysis was required for specific branches with ambiguous results.

It is implied that highly conserved regions in the dN/dS analysis are crucial in the context of pathogenicity, but genes with a dN/dS statistic >1 required further analysis via association tests. Currently, the link between these genes and phenotypes is still inconclusive.

Improvements in the pre-processing pipeline could increase confidence in downstream results. It may be worth re-sequencing isolates with low coverage to re-assess variant calls. Higher quality reads can also allow for successful phasing of the resulting alleles.

Acknowledgements

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.Bioinformatics, 28(12), 1647-1649.

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

Yang, Z. 2007 PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24, 1586-1591.