

# **Predicting Cancer Susceptibility from Single-Nucleotide Polymorphism Data: A Case Study in Multiple Myeloma**

**Michael Waddell and David Page**

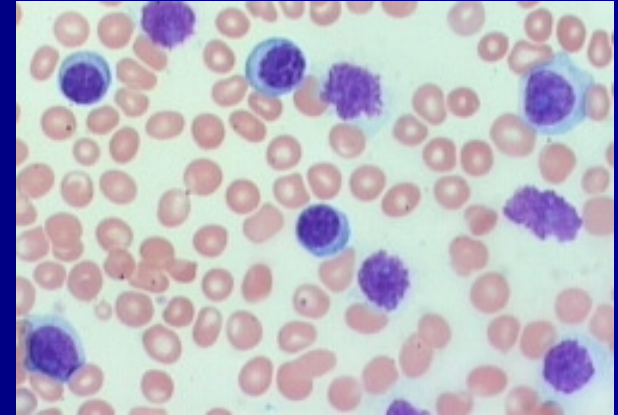
University of Wisconsin

**Fenghuang Zhan, Bart Barlogie and John Shaughnessy, Jr.**

University of Arkansas for Medical Sciences

# Multiple Myeloma

- Multiple Myeloma (MM) is a **uniformly fatal** malignancy of the plasma cells.
- Relatively **high frequency** in older adults  
(0.035% of the US population aged 70+).
- Much **lower frequency** in younger adults  
(0.002% of the US population aged 30–54).
- Can susceptibility to **early-onset** MM be **predicted** using SNP profiles?



# SNPs

- Part of the genetic variation among individuals is the **cumulative** effect of variations at a number of **single-base** locations within the genome.
- These locations are known as **SNPs** (Single Nucleotide Polymorphisms).
- A “**SNP pattern**” consists of the DNA bases present at a large number of SNP positions.

# SNPs

- SNPs can be used to identify **markers** for genes associated with **predisposition** to a disease.
- These markers provide:
  - Information about a patient's **risk** for disease
  - **Insight** into the disease process
  - Protein **targets** for novel drug therapies

# Benefits of Using SNPs

- A person's SNP pattern is highly unlikely to **change** over time or as a result of disease.
- SNP data can be collected from **any tissue** in the body (not just from diseased tissue).
- This allows a **larger number** of samples to be obtained (especially controls) since faster and **less invasive** procedures are used.

# 3 Challenges of Using SNPs

1. There are now over **one million** SNPs known but measuring them all is typically **cost-prohibitive**.
  - SNP data contain measurements for only a **small fraction** of known SNPs (typically a few thousand).
  - If **prior knowledge** is available, focus the SNPs collected to particular region(s) of the genome.
  - Otherwise, choose SNPs to give good **overall coverage** of the genome.

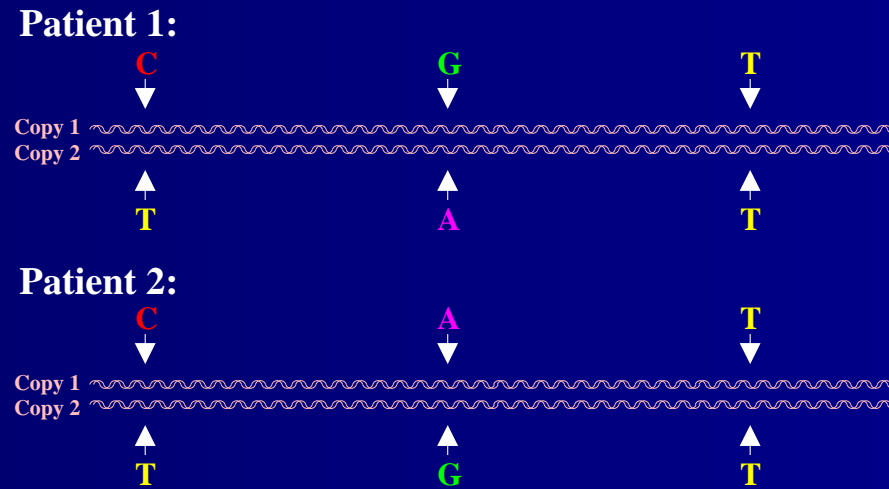
(Subsequent studies can use these results as **prior knowledge**.)

# 3 Challenges of Using SNPs

2. SNP data commonly contain **missing values**.
  - This can **adversely affect** many algorithms used for classification tasks.
  - When choosing an **algorithm** to use, this must be taken into consideration in order to choose an **appropriate** one.

# 3 Challenges of Using SNPs

## 3. SNP data are “unphased.”



	SNP 1		SNP 2		SNP 3		...	class
Patient 1	C	T	A	G	T	T	...	Diseased
Patient 2	C	T	A	G	T	T	...	Healthy
...	...	...	...	...	...	...	...	...

# 3 Challenges of Using SNPs

## 3. SNP data are “unphased.”

- 2 main ways of dealing with this:
  1. Perform **haplotyping** to determine the phasing of the SNP data.
    - Algorithms for haplotyping are **not** guaranteed to be correct.
    - These algorithms generally require data on related individuals (**pedigrees**).
  2. Work with the data in its **unphased form**.

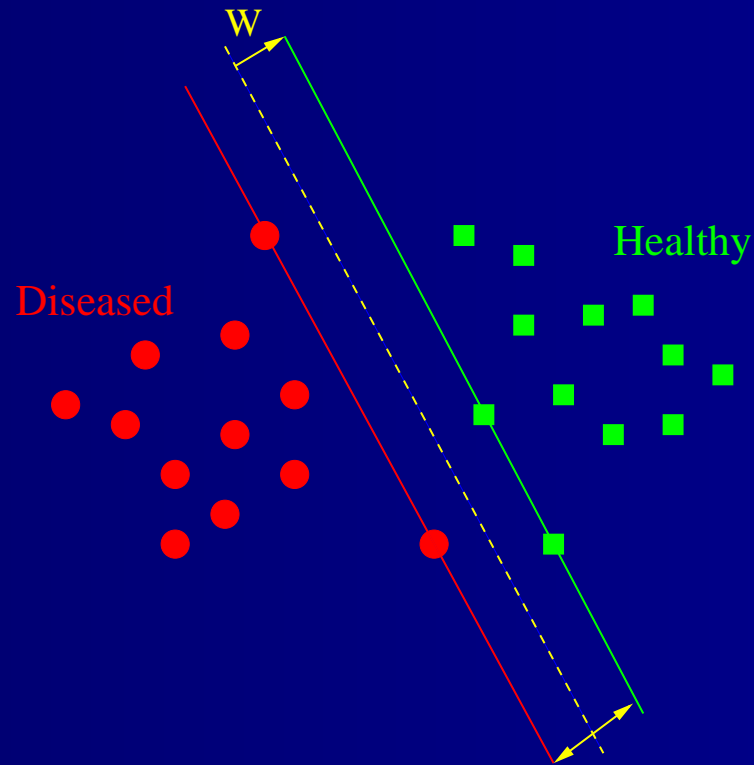
# Methodology

1. Divide the patients into **two groups**: **healthy** and **disease**.
2. Machine learning or statistical modeling algorithms can then be used to construct a **model** of the disorder.
3. This model can be validated using **cross-validation**.
4. If the model is sufficiently accurate, it can be studied to gain **insight** into the disease.

# Data Set

- “Unphased” SNP data for 80 patients (based on 3000 SNPs)
  - 40 “predisposed” patients: diagnosed with MM before age 40
  - 40 “not predisposed” patients: diagnosed with MM after age 70
- The SNPs were selected to give good overall coverage of the human genome — *not* based on prior knowledge of MM.

# Support Vector Machines



**Figure 1:** Linear support vector machines (SVMs) maximize the “margin” between the bounding planes.

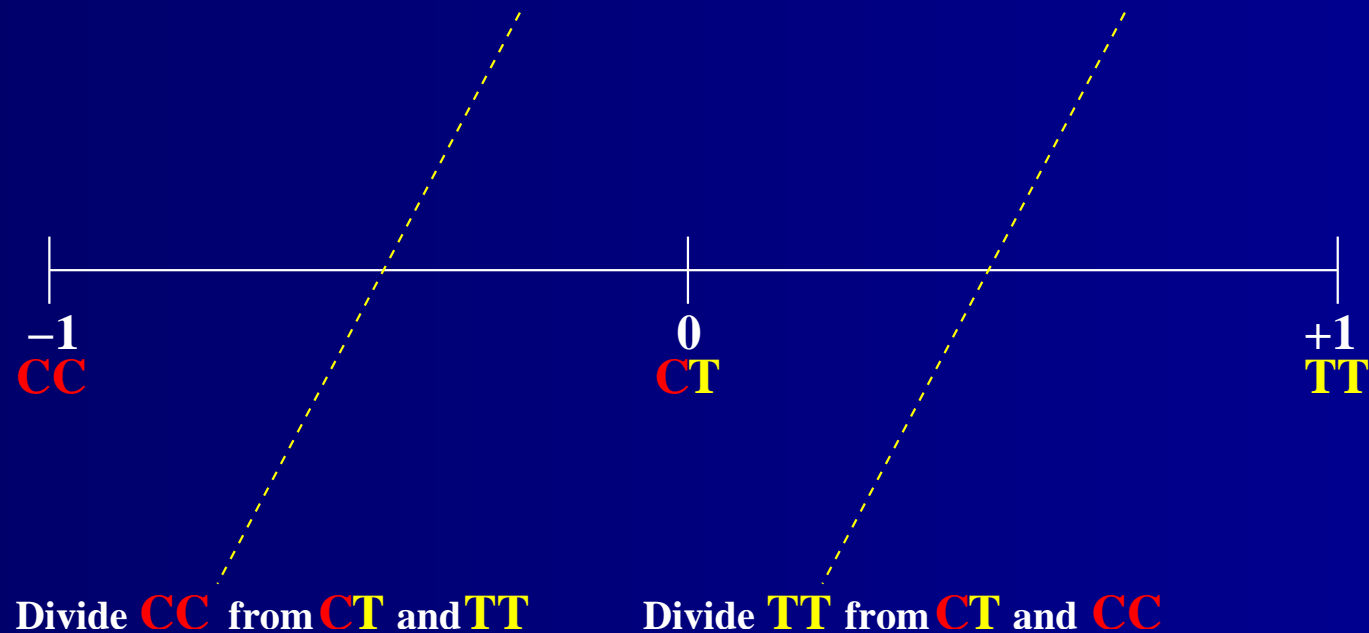
# Support Vector Machines

- Non-linear SVMs also exist, but their output is harder to gain **insights** from.
- With a linear SVM, features (SNPs) with larger **coefficients** in the linear separator are more **important**.
- SVMs (whether linear or not), require features to be **numerical** and **continuous**.
- SNP data is **not** numerical or continuous.

# SVMs Need Continuous Features

- In unphased SNP data, each feature takes on one of **3 discrete values**:
  1. Homozygous for the dominant SNP
  2. Homozygous for the recessive SNP
  3. Heterozygous
- Therefore, we convert the **discrete** values to the values: -1, 0, 1  
(Where 0 represents heterozygous and the two homozygous cases are arbitrarily mapped to  $\pm 1$ .)
- “Phased” SNP data would have a **4<sup>th</sup>** value.

# SVMs Need Continuous Features



**Figure 2:** This choice of mapping is made because it allows the SVM to divide between the presence and absence of either SNP.

# SVMs Need Continuous Features

- This mapping allows dividing between the **presence** and **absence** of either SNP.
- It does **not** allow the SVM to divide between the presence and absence of **homozygosity**.  
(But this is unlikely to be **biologically relevant**.)
- An alternative mapping to allow **both** of these divisions would use **2 features** per SNP.  
(However, this would **double** the number of features and machine learning algorithms perform better with **fewer** features.)

# "Curse of Dimensionality"

- Having **many more** features than examples
- A major **problem** in machine learning
- We employ 3 methods to deal with this:
  1. Use leave-one-out **cross-validation** to assess the accuracy of the model since it is robust to high-dimensional data.
  2. Use SVMs because they are more **robust** than some other algorithms on high-dimensional data.
  3. Use **feature selection** to reduce the number of features given to the SVM.

# Feature Selection

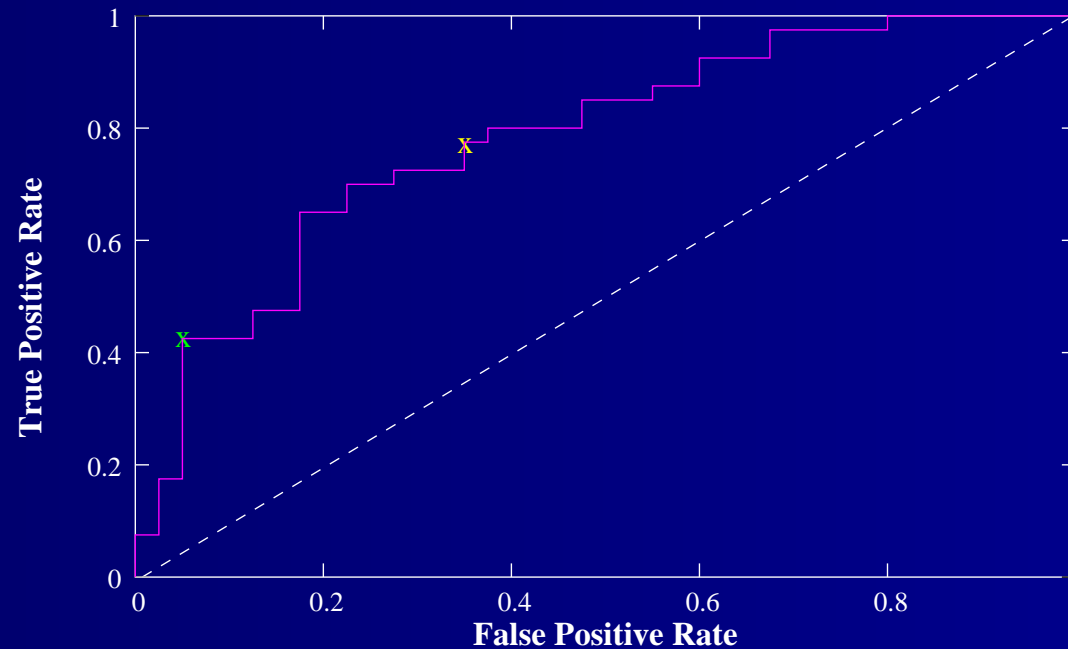
- Feature selection is commonly used to **improve performance** of modeling algorithms.
- It is common, **though incorrect**, to perform feature selection once over the whole data set.
- This practice leads to information **“leaking”** from the validation set into the training set.
- To avoid this, the top 10% of SNPs were chosen by information gain **on each fold** of cross-validation prior to model construction.

# Results

		Predicted	
		Not Predisposed	Predisposed
Actual	Not Predisposed	31	9
	Predisposed	14	26

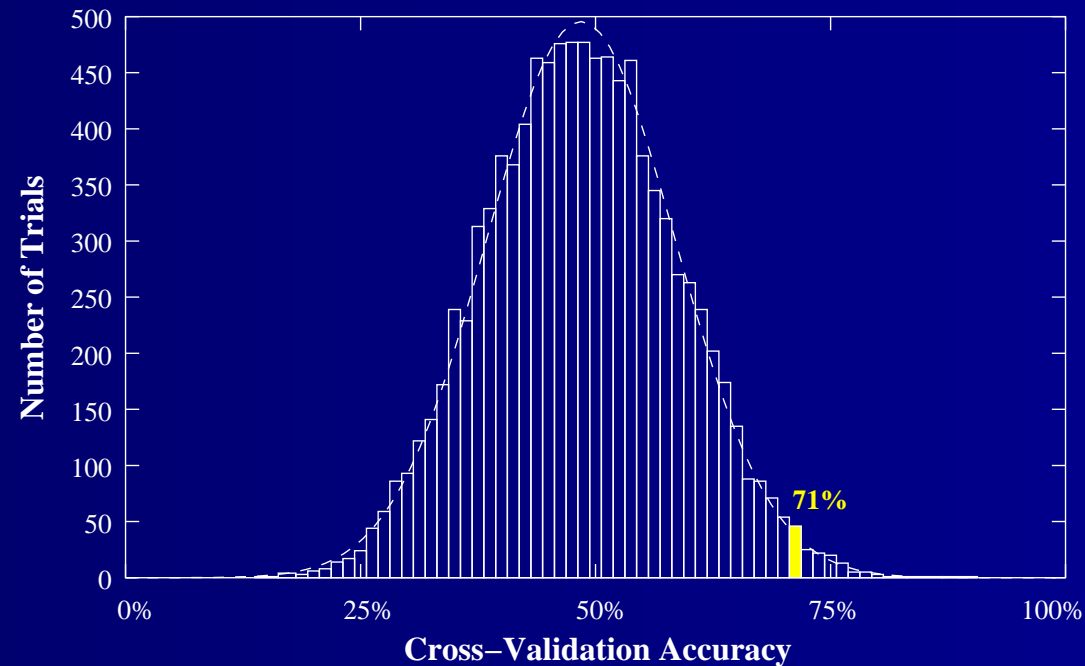
**Figure 3:** Our approach yields an accuracy estimate of **71%** by leave-one-out cross validation. This is **significantly** better than random guessing ( $p < 0.01$ ).

# ROC Curve



**Figure 4:** The Receiver-Operator Characteristic (ROC) curve shows that linear SVMs perform **significantly** better than random guessing (dotted line). It also shows the accuracy if we tuned the SVM model to **bound** the false positive rate (since MM is rare). The point (5%, 42.5%) is noted in **green**. The point without tuning (35%, 77.5%) is noted in **yellow**.

# Permutation Test



**Figure 5:** Permutation testing assesses **dependency** of the classifier to that particular data set. Provides an estimate of the error of the classifier when data is limited. Our result of 71% is **significant** at the  $p < 0.05$  level (2-tailed). This agrees with the results of a standard **binomial test** (also performed).

# Control Data

- We want to show that the model is **not** based on the **age** of the patients
- 2911 SNPs were obtained on 28 unrelated persons **without MM**
  - 14 were **older** than 70 years-of-age
  - 14 were **younger** than 40 years-of-age
- Again, SNPs were chosen for **broad coverage** of the genome
- Used **same methodology** as before

# Control Results

		Predicted	
		Over 70	Under 40
Actual	Over 70	6	8
	Under 40	7	7

**Figure 6:** Control accuracy is **46%** by leave-one-out cross validation. Although the 2911 SNPs were a different set than the original 3000, this result makes it **unlikely** that our original 71% accuracy is due to **predicting age**.

# Conclusions

- Our accuracy of **71%** is not as high as we have obtained using microarray data.
- However, this prediction is based **only** on SNP data (which are not affected by disease progression like microarray data).
- Also, our SNP coverage was **relatively sparse** (only 3000 SNPs were used).
- Thus, we conclude that SNP data **do** provide predictive ability for cancer susceptibility.

# Future Work

- Expand the study to include **more patients** in order to further validate these results (including **more controls**).
- Use a **denser coverage** of the genome than the 3000 SNPs used.
- Use a **more focused** coverage of the genome: focused on those regions found in this study to be significant for MM predisposition.

# Future Work

- Further tune the SVM algorithm to use a **smaller set** of features for classification to gain better **insight** into those regions/genes that are important.
- Compare the SNPs found to be important with the genes found using related **gene microarray** studies.

# Acknowledgments

- **Computation and Informatics in Biology and Medicine (CIBM) Training Program**
- **National Library of Medicine**  
Grant No. 5T15LM007359
- **SVM<sup>light</sup>** (<http://svmlight.joachims.org>)

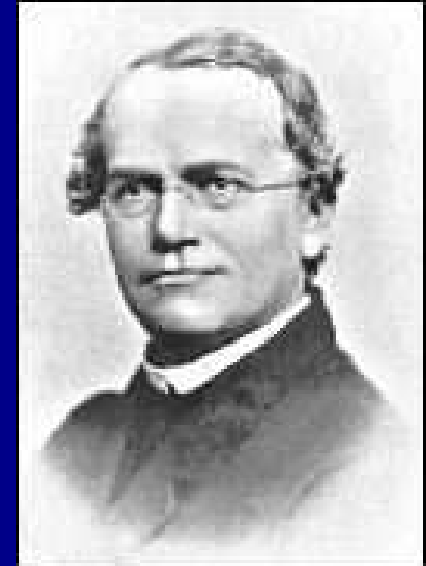
*Thank you for your attention!*

# Genetics Background

- Genetics technology has **grown** dramatically in the past 2 decades.
  - PCR technology developed (1985)
  - BLAST algorithm developed (1990)
  - DNA chips made commercially available (1996)
  - SNP consortium founded (1999)
  - Human Genome Project working draft published (2001)
- This growth has enabled **identification** of genes responsible for predisposition to some **inherited** human disorders.

# Gregor Mendel (1822–1884)

- However, these successes have been primarily limited to simple **Mendelian** disorders.
- Mendelian (or **monogenic**) disorders are rare.
- The vast majority of disorders are likely **polygenic**.



# Finding genes

- There are two main approaches for finding genes responsible for disorders:
  1. Linkage analysis and positional cloning
  2. Direct analysis of candidate genes (via association studies)

# Positional Cloning

- Requires **families** with known pedigrees.
- Genetic markers (RFLPs, microsatellite loci, SNPs, etc.) closest to the gene(s) of interest will be strongly **correlated** with the disease pattern in the family.
- Results may not be **generalizable**.

# Association Studies

- Also uses genetic **markers**
- Does **not** require families
- Lots of **false positives** (esp. with non-monogenic disorders)
  - More features than datapoints
  - Low prior probability of association
  - Can be misled by **population stratification** (ie: ethnicity, gender, etc.)

# Finding genes

- These standard approaches have **successfully** identified genes causing:
  - Breast cancer (BRCA-1 & -2)
  - Colon cancer (FAP & HNPCC)
  - Diabetes (MODY-1, -2 & -3)
  - etc.  
(All of which are Mendelian or near-Mendelian.)
- These standard approaches **fail** when attempting to identify a set of genes, each of **modest** effect, whose combined effects cause a particular trait (aka: a **polygenic** trait).

# Genetic Heterogeneity

- Distinct mutations that cause the same, **indistinguishable**, phenotype.
- Types of genetic heterogeneity:
  1. Distinct loci that **interact** to cause the phenotype
  2. Distinct loci that are capable of **independently** causing the phenotype
- Standard techniques can deal with **small** numbers of independent loci.
- They fail with **large** numbers of independent loci or **interactions** among loci.

# SNPs Are Stable Over Time

- SNPs have been shown to be extremely stable over **evolutionary time**
- SNPs are not as susceptible to **age-related mutations** as microsatellite polymorphisms
- We used DNA from peripheral blood **mononuclear cells**.
  - This mixture of cells should **not** have an over-representation of any given clone containing a **specific** mutation.

# Discriminating on Homozygosity

- If heterozygosity does **not** predisposes cancer but either homozygous state **does**:
  - Regardless of the relative abundance, a very **large** percentage of the population would be **homozygous** for one variant or the other
  - Thus, this feature would **not** be informative and would be **ignored** by our model
  - Also, this would lead to a much **higher prevalence** of MM than we see in practice

# Discriminating on Homozygosity

- If heterozygosity **does** predisposes cancer but either homozygous state does **not**:
  - If both allelic variants were **common**, this feature would be **uninformative** and ignored
  - If one variant was **rare**, then a homozygote of that variant would be **very** rare and would thus **not** affect the model significantly

# Interpreting SVM Results

- Ideally, the SVM model would be based on only a **few** SNPs.
- An SVM only “uses” those SNPs with **non-zero coefficients**.
- Our SVM produced a model that uses over **150** SNPs.
- 48 SNPs had non-zero coefficients on **every** cross-validation fold.

# SNPs With Non-Zero Coefficients

SNP	Chrom.	Contig Accession	Contig Position	Chrom. Position	Hit Orientation
<i>TSC0022568</i>	20	NT_011387.8	23917072	23925072	minus strand
<i>974272</i>	8	NT_008046.11	6451653	97542762	minus strand
<i>930891</i>	5	NT_023132.10	2310237	171083062	plus strand
<i>930311</i>	2	NT_037537.1	1516282	172969092	plus strand
<i>921897</i>	5	NT_006431.11	4641524	61626387	plus strand
<i>912797</i>	1	NT_004636.13	1545773	66344816	plus strand
<i>910069</i>	6	NT_025741.10	16628671	139745590	plus strand
<i>880307</i>	3	NT_022509.9	2184897	179348004	minus strand
<i>869380</i>	1	NT_004612.13	1632279	210865199	plus strand
<i>759002</i>	7	NT_033968.2	1421567	51360645	plus strand
<i>755614</i>	3	NT_037588.0	31998	unplaced	minus strand
<i>737453</i>	2	NT_005403.11	16204291	214479968	plus strand
<i>726828</i>	15	NT_033268.0	126217	unplaced	minus strand
<i>726828</i>	3	NT_005684.9	71977	77618043	plus strand
<i>717480</i>	12	NT_035247.0	106169	unplaced	plus strand
<i>717480</i>	13	NT_009952.11	17735730	99066060	plus strand
<i>712269</i>	17	NT_030843.4	770495	19404437	minus strand
<i>707860</i>	6	NT_007592.11	8361187	17560475	plus strand
<i>679206</i>	10	NT_030059.8	8262073	102231090	plus strand

# SNPs With Non-Zero Coefficients

SNP	Chrom.	Contig Accession	Contig Position	Chrom. Position	Hit Orientation
598985	8	NT_034927.0	61452	unplaced	minus strand
598985	8	NT_023744.11	1619368	2954228	plus strand
257769	5	NT_006576.11	12192025	16122034	plus strand
220872	11	NT_033899.3	4804715	116778426	plus strand
1992291	2	NT_005204.11	6319962	27596897	plus strand
1990272	4	NT_006342.13	974650	10545064	plus strand
1989229	17	NT_037978.1	98150	unplaced	plus strand
1989229	17	NT_010748.10	88877	45477210	minus strand
1882055	7	NT_007819.11	35393849	35742551	minus strand
1863239	7	NT_007819.11	36884313	37233015	minus strand
1862336	19	NT_011109.13	9443320	47679528	plus strand
1848280	2	NT_005265.11	8167756	188270833	minus strand
1544035	18	NT_025028.11	4990909	68576178	plus strand
1540801	3	NT_022517.13	1185663	19978681	plus strand
1519822	8	NT_023679.13	2113240	50269239	minus strand
1510332	3	NT_022463.12	879640	162748670	plus strand
1458904	8	NT_008251.11	403353	34579397	plus strand
1411426	9	NT_035014.2	1022925	127413028	minus strand
1396614	5	NT_006713.11	6482883	77351542	minus strand

# SNPs With Non-Zero Coefficients

SNP	Chrom.	Contig Accession	Contig Position	Chrom. Position	Hit Orientation
1371374	5	NT_016755.11	2785127	124857651	plus strand
1366199	5	NT_030685.4	553283	115729193	plus strand
1335286	13	NT_024524.11	27334658	74268506	plus strand
1332234	9	NT_023974.13	3907284	31122790	plus strand
1275054	11	NT_033899.3	13114602	125088313	minus strand
124756	7	NT_007933.10	52899826	126200901	plus strand
116016	14	NT_026437.9	34866829	65373186	minus strand
1028484	6	NT_007540.11	2825315	64980073	plus strand
1024892	2	NT_022184.10	1857674	71585093	minus strand
1023617	5	NT_023132.10	4669166	173441991	plus strand
1019791	5	NT_034779.2	4954101	155150577	minus strand
1017935	4	NT_022782.10	2048357	42629852	minus strand
1014025	17	NT_010799.11	881757	28044553	plus strand
1002624	14	NT_037845.1	13318680	26562678	plus strand