

*Modeling Patterns in Single-Nucleotide  
Polymorphism Data for Predicting Cancer  
Susceptibility: A Case Study in Multiple Myeloma*

**Michael J. Waddell**

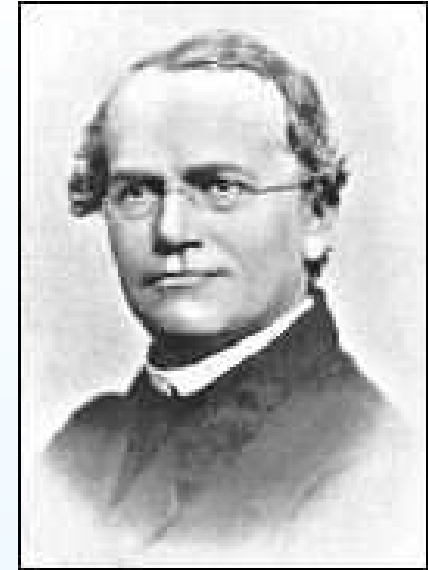
Laboratory of Dr. C. David Page  
Departments of Computer Science and  
Biostatistics and Medical Informatics  
University of Wisconsin

## 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**)

## Finding genes

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

## 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**)

## 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**)
    - ♠ 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

- 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**)
- These standard approaches have **successfully** identified genes causing:
  - ♠ Breast cancer (BRCA-1 & -2)
  - ♠ Colon cancer (FAP & HNPCC)
  - ♠ Diabetes (MODY-1, -2 & -3)
  - ♠ etc.

## 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**)
- 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.)

## 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**)
- 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*

---

- Genetic heterogeneity: distinct mutations 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.

# *Machine Learning to the Rescue*

---

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.
- ★ Algorithms have been developed that deal well with **interactions** and **redundant** features.

## 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 can be used in linkage analysis or association studies to identify **markers** for genes associated with a disorder.

## *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*

---

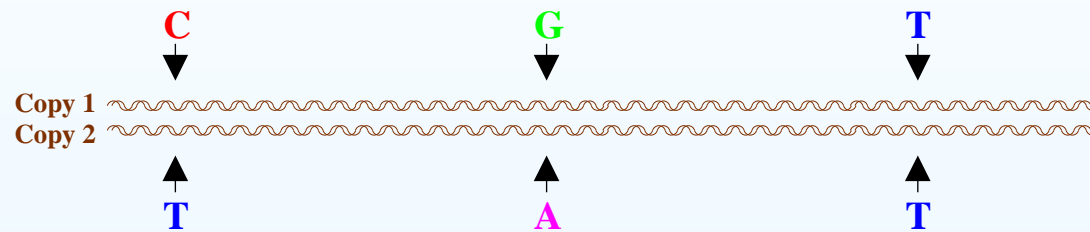
2. SNP data commonly contain **missing values**.

- 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.

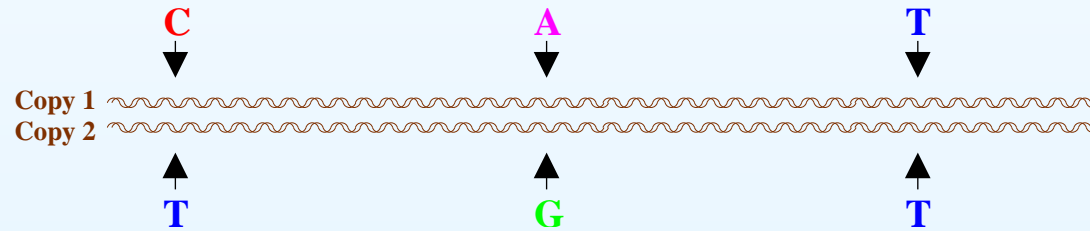
# 3 Challenges of Using SNPs

## 3. SNP data are “unphased.”

**Patient 1:**



**Patient 2:**



	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*

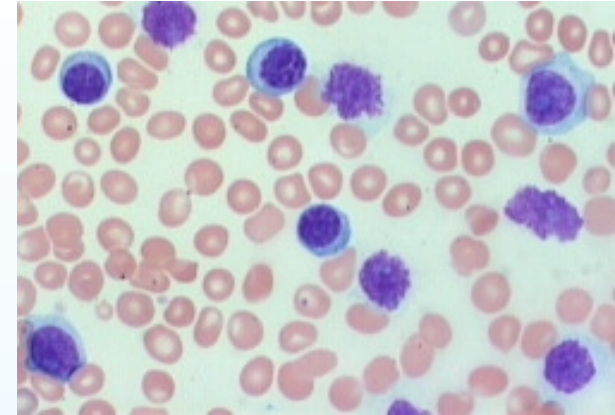
---

### 3. SNP data are “unphased.”

- There are 2 main ways of dealing with SNP data:
  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**.

# Multiple Myeloma

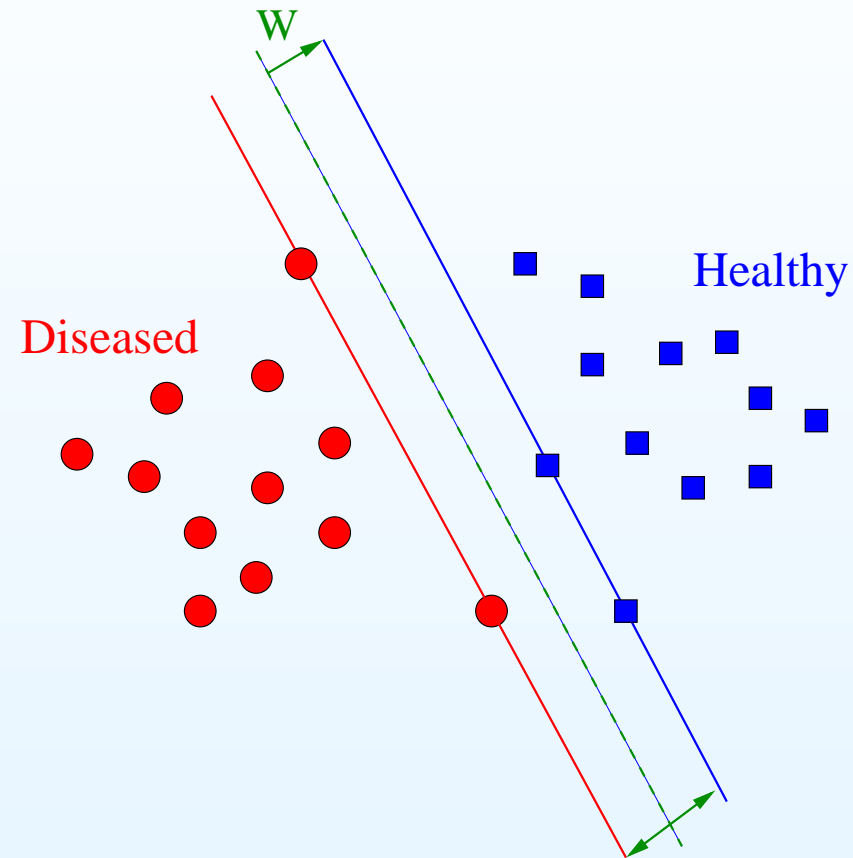
- Multiple Myeloma (MM) is a **uniformly fatal** malignancy of the plasma cells.
- MM occurs with relatively **high frequency** in older adults (0.035% of the US population aged 70+).
- MM occurs with much **lower frequency** in younger adults (0.002% of the US population aged 30–54).
- We hypothesize that those diagnosed with MM at a young age have a **genetic predisposition** to the disease.



## Data Set

- “Unphased” SNP data for 80 patients (based on 3000 SNPs)
  - ♠ 40 “old” patients: diagnosed with MM after age 70
  - ♠ 40 “young” patients: diagnosed with MM before age 40
- 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*

---

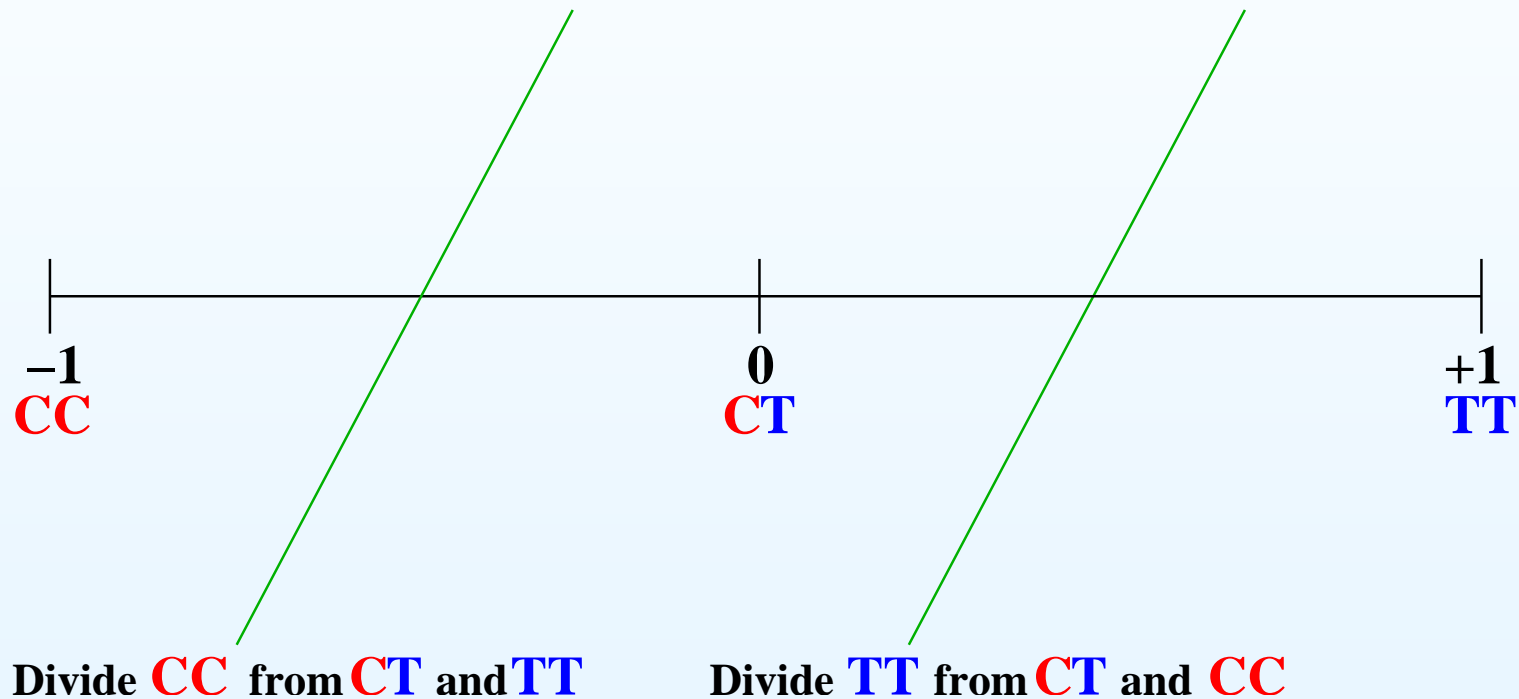
- Non-linear SVMs also exist, but their output is harder to gain **insights** from.
- With a linear SVM, features (SNPs) with larger **coefficients** in the linear separator are more **important**.
- SVMs (whether linear or not), require features to be **numerical** and **continuous**.
- SNP data is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

## *SVMs Require 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>** possible value.

# *SVMs Require 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 Require Continuous Features*

---

- This mapping allows the SVM to divide 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 that would 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"

---

- The “Curse of Dimensionality” — having many more features than examples — is 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 practice, **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 common pitfall, the top 10% of SNPs were chosen by information gain **on each fold** of cross-validation prior to model construction.

# Results

		Predicted	
		Old	Young
Actual	Old	31	9
	Young	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 ( $p < 0.01$ ).

## 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.

## *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
TSC0022568	20	NT_011387.8	23917072	23925072	minus strand
974272	8	NT_008046.11	6451653	97542762	minus strand
930891	5	NT_023132.10	2310237	171083062	plus strand
930311	2	NT_037537.1	1516282	172969092	plus strand
921897	5	NT_006431.11	4641524	61626387	plus strand
912797	1	NT_004636.13	1545773	66344816	plus strand
910069	6	NT_025741.10	16628671	139745590	plus strand
880307	3	NT_022509.9	2184897	179348004	minus strand
869380	1	NT_004612.13	1632279	210865199	plus strand
759002	7	NT_033968.2	1421567	51360645	plus strand
755614	3	NT_037588.0	31998	unplaced	minus strand
737453	2	NT_005403.11	16204291	214479968	plus strand
726828	15	NT_033268.0	126217	unplaced	minus strand
726828	3	NT_005684.9	71977	77618043	plus strand
717480	12	NT_035247.0	106169	unplaced	plus strand
717480	13	NT_009952.11	17735730	99066060	plus strand
712269	17	NT_030843.4	770495	19404437	minus strand
707860	6	NT_007592.11	8361187	17560475	plus strand
679206	10	NT_030059.8	8262073	102231090	plus strand
598985	8	NT_034927.0	61452	unplaced	minus strand
598985	8	NT_023744.11	1619368	2954228	plus strand

# SNPs With Non-Zero Coefficients

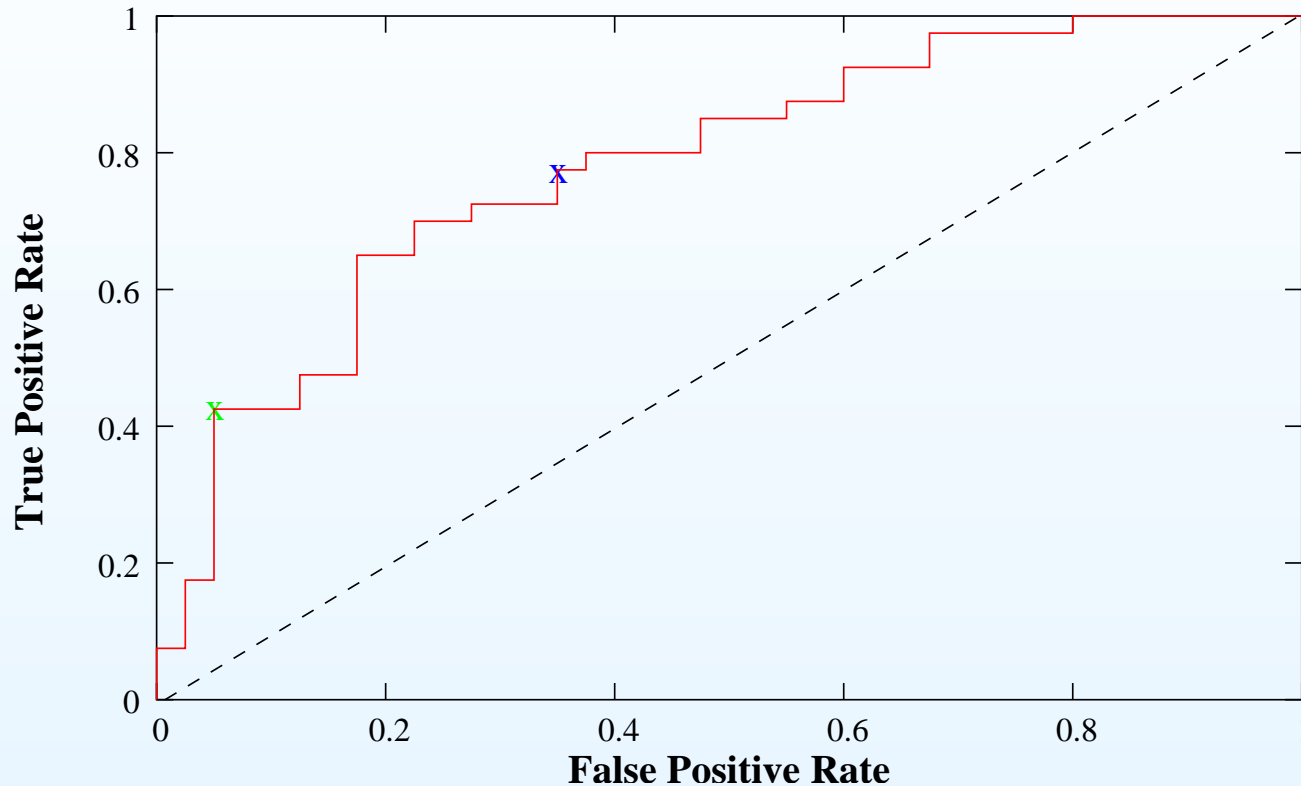
SNP	Chrom.	Contig Accession	Contig Position	Chrom. Position	Hit Orientation
257769	5	NT_006576.11	12192025	16122034	plus strand
220872	11	NT_033899.3	4804715	116778426	plus strand
1992291	2	NT_005204.11	6319962	27596897	plus strand
1990272	4	NT_006342.13	974650	10545064	plus strand
1989229	17	NT_037978.1	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
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

# *SNPs With Non-Zero Coefficients*

---

<b>SNP</b>	<b>Chrom.</b>	<b>Contig Accession</b>	<b>Contig Position</b>	<b>Chrom. Position</b>	<b>Hit Orientation</b>
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

# Receiver Operator Characteristic (ROC) Curve



**Figure 4:** The 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 **blue**.

## *Comparisons with Other Algorithms*

---

- After finishing analysis of the linear SVM results, we decided to try some **other algorithms** on this problem:
  - ♠ non-linear (Gaussian and polynomial) SVMs
  - ♠ decision trees
  - ♠ naïve Bayesian networks
  - ♠ ensembles of voting decision stumps

## *Comparisons with Other Algorithms*

<b>Algorithm</b>	<b>“Old” Accuracy</b>	<b>“Young” Accuracy</b>	<b>Overall Accuracy</b>
Linear SVMs	65.0%	77.5%	71.2%
Polynomial SVMs	50.0%	70.0%	60.0%
Gaussian SVMs	0.0%	0.0%	0.0%
Decision Trees	45.0%	52.5%	48.8%
Boosted Decision Trees	50.0%	50.0%	50.0%
Naïve Bayes	42.5%	45.0%	43.8%
Ensemble of Voters	10.0%	40.0%	25.0%

## *Comparisons with Other Algorithms*

---

- We see from this comparison, that our choice of using **linear SVMs** for this task was good.
- However, this comparison raises a number of questions:
  1. Why did **polynomial** SVMs do worse than linear?
  2. Why did **Gaussian** SVMs get 0% accuracy?
  3. Why did all other algorithms do the **same or worse** than random guessing?

## *Comparisons with Other Algorithms*

- We see from this comparison, that our choice of using **linear SVMs** for this task was good.
- However, this comparison raises a number of questions:
  1. Why did **polynomial** SVMs do worse than linear?
    - ♠ Polynomial SVMs **can** separate between the absence and presence of homozygosity (which is not biologically relevant) they were likely **led astray** by irrelevant correlations.
  2. Why did **Gaussian** SVMs get 0% accuracy?
  3. Why did all other algorithms do the **same or worse** than random guessing?

## *Comparisons with Other Algorithms*

---

- We see from this comparison, that our choice of using **linear SVMs** for this task was good.
- However, this comparison raises a number of questions:
  1. Why did **polynomial** SVMs do worse than linear?
  2. Why did **Gaussian** SVMs get 0% accuracy?
    - ♠ Because of the very **large** number of features compared to the number of patients, it is possible that Gaussian SVMs fit the training data so tightly that it simply **memorized** that data and was not able to generalize it at all.
    - ♠ Other **ideas**?
  3. Why did all other algorithms do the **same or worse** than random guessing?

## *Comparisons with Other Algorithms*

- We see from this comparison, that our choice of using **linear SVMs** for this task was good.
- However, this comparison raises a number of questions:
  1. Why did **polynomial** SVMs do worse than linear?
  2. Why did **Gaussian** SVMs get 0% accuracy?
  3. Why did all other algorithms do the **same or worse** than random guessing?
    - ♠ Naïve Bayes and EOVs assume **feature independence** which is strongly violated in this domain.
    - ♠ Decision trees are not only able to separate presence and absence of homozygosity, but they are **not robust** with high-dimensional data.

## *Future Work*

---

- Expand the study to include **more patients** in order to further validate these results.
- 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.
- 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.

## Future Work

- Obtain **control** data points — SNP data on individuals, at both “young” and “old” ages, without MM — to **validate** that SNP patterns do not change with age.
- Compare the SNPs found to be important with the genes found using related **gene microarray** studies.

# *Acknowledgments*

---

- **David Page**  
University of Wisconsin
- **Fenghuang Zhan, Bart Barlogie and John Shaughnessy, Jr.**  
University of Arkansas for Medical Sciences
- **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>)
- **C5.0** (<http://www.rulequest.com>)
- **EOV** (<http://www.cs.wisc.edu/~mwaddell/eov.html>)