

Project Notes:

Project Title: Implementation of the Graphlet Screening Method in Genomic Analysis Using Hail

Name: Rishit Avadhuta

Contents:

Knowledge Gaps:	2
Literature Search Parameters:	3
Tags:	3
Template Title	5
Article #1 Notes: RNA-Seq: Basics, Applications, and Protocol	6
Article #2 Notes: RNA-Seq: a revolutionary tool for transcriptomics	8
Article #3 Notes: Advantages of paired-end and single-read sequencing	12
Article #4 Notes: GWAS Fact Sheet (NIH)	14
Article #5 Notes: Optimality of Graphlet Screening in High Dimensional Variable Selection	16
Article #6 Notes: What is a Linear Regression Model?	18
Article #7: Understanding Ordinary Least Squares (OLS) Regression	20
Article #8: Genomic regression analysis of coordinated expression	22
Article #9: Variable selection and estimation in high-dimensional models	25
Article #10: Efficient Regularized Regression with L_0 Penalty for Variable Selection and Network Construction	27
Patent #1: Artificial Intelligence-Based Sequencing	29
Patent #2: Deep learning-based techniques for pre-training deep convolutional neural networks	31
Article #11: LASSO regression	32
Article #12: Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods	33
Article #13: Overview of Statistical Methods for Genome-Wide Association Studies (GWAS)	35
Article #14: Polymorphic Regions Affecting Human Height Also Control Stature in Cattle	37
Article #15: Variable selection in linear regression models: Choosing the best subset is not always the best choice	39
Article #16: Meta-regression of genome-wide association studies to estimate age-varying genetic effects	41

Article #17: High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in Genome Wide Association Studies	43
Article #18: An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies	45
Article #19: Detecting Differentially Expressed Genes with RNA-seq Data Using Backward Selection to Account for the Effects of Relevant Covariates	47
Article #20: Regularization and Variable Selection Via the Elastic Net	49

Knowledge Gaps:

This list provides a brief overview of the major knowledge gaps for this project, how they were resolved and where to find the information.

Knowledge Gap	Resolved By	Information is located	Date resolved
What is a GWAS?	Research – NIH Article	https://www.genome.gov/	08/15/2024
What is Graphlet Screening?	Research Paper – Optimality of Graphlet Screening in High Dimensional Variable Selection	https://www.jmlr.org/papers/v15/jin14a.html	9/19/2024
What is RNA-Seq?	Research – Articles #1, #2, #3	Articles #1, #2, #3	8/11/2024
What is Linear Regression and what methods exist to do it?	Research – Articles #6, #7, #9, #10, and Patent #2	Articles #6, #7, #9, #10, and Patent #2	10/10/2024
How can linear regression be used in genomics?	Research – Article #8 and Patent #1	Article #8 and Patent #1	10/10/2024

Literature Search Parameters:

These searches were performed between 08/11/2024 and ...

List of keywords and databases used during this project.

Database/search engine	Keywords	Summary of search
Google / Google Scholar	Rna-seq, what is rna-seq,	General websites, some long journals which detailed on methodology
Google	GWAS, what is GWAS, why is GWAS important	Medical sources and genomic websites
Google	Linear regression	Algorithms used to assemble linear regression models
Google	Ordinary Least Squares	The OLS regression method and how it is used to estimate beta correlation
Google Scholar	High dimensional variable selection	Different regression methods for determining the correlation between a response vector and a design matrix
Google Scholar	L_0 -Penalization	Purpose of L_0 -penalization itself, papers optimizing the model, and comparisons to other models
Google Patents and USPTO Patent Search	Linear Regression	Patents displaying the different applicative uses for linear regression

Tags:

Tag Name	
#rna #rnaSeq	#dna
#sequencing	#genomics

#methodology	#genomics
#methods	#lrRNAseq and #srRNAseq
#GWAS	#FactSheet
#NIH	#NHGRI
#statistics #stats #statsmodel	#algorithm
#linearRegression #regression	#matlab #R

Template Title

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	
Source citation (APA Format)	
Original URL	
Source type	
Keywords	
#Tags	
Summary of key points + notes (include methodology)	
Research Question/Problem/ Need	
Important Figures	
VOCAB: (w/definition)	
Cited references to follow up on	
Follow up Questions	

Article #1 Notes: RNA-Seq: Basics, Applications, and Protocol

Source Title	RNA-Seq: Basics, Applications, and Protocol
Source citation (APA Format)	Mackenzie, R. J. (2018, April 6). RNA-seq: Basics, applications and protocol. <i>Technology Networks Genomics Research</i> . https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461
Original URL	https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461
Source type	Newspaper Article (Online)
Keywords	RNA-Seq, Sequencing, Genomics, Methodology
#Tags	#rna #dna #sequencing #genomics #methodology
Summary of key points + notes (include methodology)	RNA Sequencing or RNA-Seq is a type of experiment which uses next generation sequencing (NGS) to analyze the transcriptome, which is the total sum of messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). This experiment aims to provide more information on diseases in which the function of genes plays a key role. When RNASeq and NGS are used together, they can help to measure transcription levels, repression, gene expression, and more. RNA-Seq involves extracting RNA, reverse transcription into cDNA, amplifying the RNA sample(s), and data analysis to determine final conclusions. Depending on what sort of RNA-Seq experiment one could pursue, they should consider changing different factors of the experiment, such as choosing between single-end or paired-end sequencing, read depth, platform choice, and more. Balancing the amount of reads with sequencing depth and isolating the mRNA (which is crucial to gene expression) are some challenges that are common in RNA-Seq experiments.
Research Question/Problem/Need	What is RNA/Next Generation-Sequencing?
Important Figures	 <p>The figure illustrates the RNA-seq workflow. On the left, a diagram shows pre-mRNA with exons and introns, which is processed into mature mRNA. Short reads are generated from the mRNA. A note states: 'Short read is split by intron when aligning to reference Genome'. The main workflow diagram shows: 'Samples of interest' (e.g., Grafton, Grafton 2) leading to 'Isolate RNA', then 'Generate cDNA, fragment, size select, add index', followed by 'Sequencing reads' (100s of millions of paired reads, 10s of billions bases of sequence). The final step is 'Downstream analysis', which includes 'Map to Genome, transcripts and predicted exon junctions', 'Stranded RNA', 'RNA-seq', 'Start read', and 'Short read split by intron'.</p> <p>Figure 1. RNA-seq data uses short reads of mRNA which is free of intronic non-coding DNA. These reads must then be aligned back to the reference genome.</p>

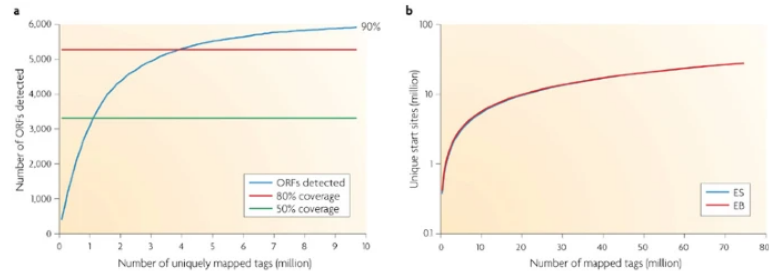
	<p><i>Credit: Technology Networks</i></p> <p>Figure 2. A workflow for RNA-seq. <i>Credit: Technology Networks</i></p>
VOCAB: (w/definition)	<p>Transcriptome – the set of all types of RNA transcripts (mRNA, rRNA, tRNA)</p> <p>Transcription – The process of turning DNA into RNA</p> <p>Translation – The process of turning RNA into proteins</p>
Cited references to follow up on	<p>https://www.technologynetworks.com/genomics/articles/transcription-vs-translation-worksheet-323080</p>
Follow up Questions	<p>What’s the difference between short-read and long-read RNA-Seq?</p> <p>How long do these experiments usually take?</p> <p>What type of RNA is the most useful to sequence?</p>

Article #2 Notes: RNA-Seq: a revolutionary tool for transcriptomics

Source Title	RNA-Seq: a revolutionary tool for transcriptomics
Source citation (APA Format)	Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. <i>Nature Reviews. Genetics</i> , 10(1), 57–63. https://doi.org/10.1038/nrg2484
Original URL	https://www.nature.com/articles/nrg2484
Source type	Journal Article
Keywords	RNA-Seq, Genomics, Transcriptomics, Methodology,
#Tags	#rna #transcriptomics #genomics #methods #methodology
Summary of key points + notes (include methodology)	This scientific publication / collaborative study focuses on long-read RNA-Sequencing (lrRNA-seq), various methods of approach, and which ones are the most effective. lrRNA-seq involves looking at a larger genetic sample than that of short-read RNA Sequencing, without breaking it into smaller fragments. This tends to return more accurate results than short-read RNA Sequencing for specific genetic variations. The study had a few main findings. Firstly, increased read quantity does not always lead to more accurate transcripts. Additionally, the study found that read quality and length led to more accurate transcripts, finding that the studies done with the technologies that prioritized these traits to be more accurate. The analysis tools they chose also affected the results, with some favoring more known transcripts and some favoring rarer transcripts. Different factors like quantification or the length of the data or what machinery was used played a noteworthy role in the accuracy of the data.
Research Question/Problem/Need	How do different methods of approach through analysis and RNA-isolation change the results from an RNA-Sequencing experiment?

Important Figures

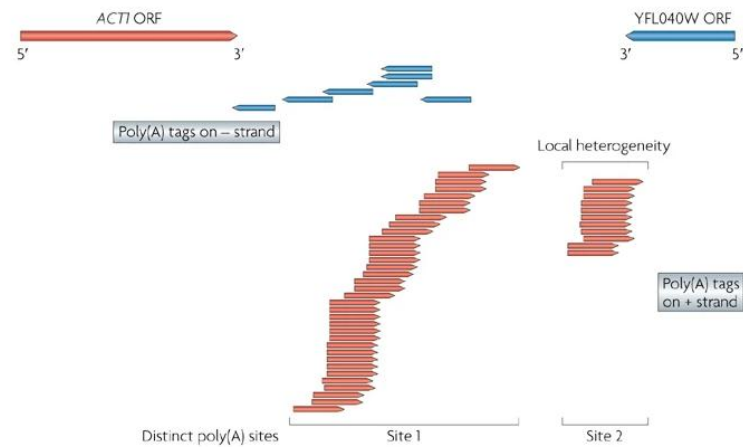
Figure 5: Coverage versus depth.



Nature Reviews | Genetics

a | 80% of yeast genes were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3' end. Data is taken from Ref. 18. **b** | The number of unique start sites detected starts to reach a plateau when the depth of sequencing reaches 80 million in two mouse transcriptomes. ES, embryonic stem cells; EB, embryonic body. Figure is modified, with permission, from Ref. 22 © (2008) Macmillan Publishers Ltd. All rights reserved.

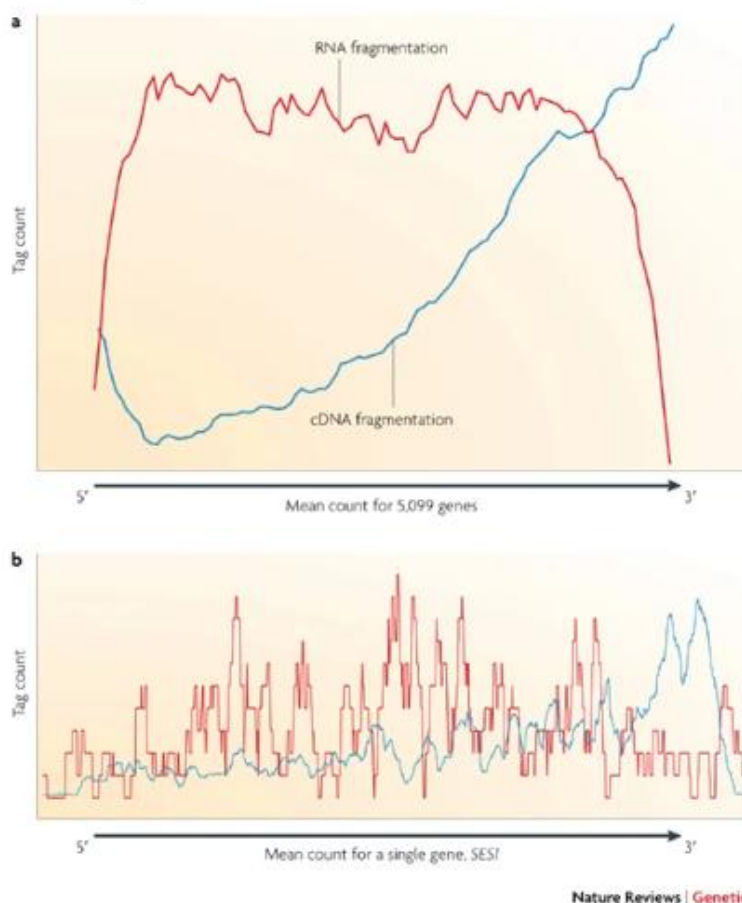
Figure 4: Poly(A) tags from RNA-Seq.



Nature Reviews | Genetics

A region containing two overlapping transcripts (*ACT1*, from the actin gene, and YFL040W, an uncharacterized ORF) from the *Saccharomyces cerevisiae* genome is shown. Arrows point to transcription direction. The poly(A) tags from RNA-Seq experiments are shown below these transcripts, with arrows indicating transcription direction. The precise location of each locus identified by poly(A) tags reveals the heterogeneity in poly(A) sites, for example, *ACT1* has two big clusters, both with a few bases of local heterogeneity. The transcription direction revealed by poly(A) tags also helps to resolve 3'-end overlapping transcribed regions¹⁸.

Figure 3: DNA library preparation: RNA fragmentation and DNA fragmentation compared.




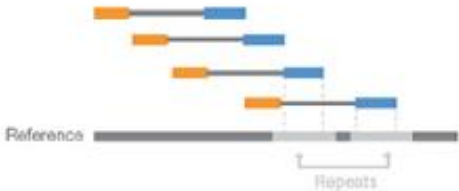
a | Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3' end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5' and 3' ends. Note that the ratio between the maximum and minimum expression level (or the dynamic range) for microarrays is 44, for RNA-Seq it is 9,560. The tag count is the average sequencing coverage for 5,000 yeast ORFs¹⁸. **b** | A specific yeast gene, *SES1* (seryl-tRNA synthetase), is shown.

<p>VOCAB: (w/definition)</p>	<p>Transcriptomics – The study of the transcriptome in regards to big data Hybridization – When fluorescently labelled cDNA is incubated with custom-made microarrays to deduce and quantify the transcriptome. Microarray – Tool used to determine whether DNA is mutated</p>
<p>Cited references to follow up on</p>	<p>https://doi.org/10.1126%2Fscience.1069415 https://doi.org/10.1126%2Fscience.270.5235.484</p>
<p>Follow up Questions</p>	<ol style="list-style-type: none"> 1. What about short reads? 2. Why are some analysis tools better than others?

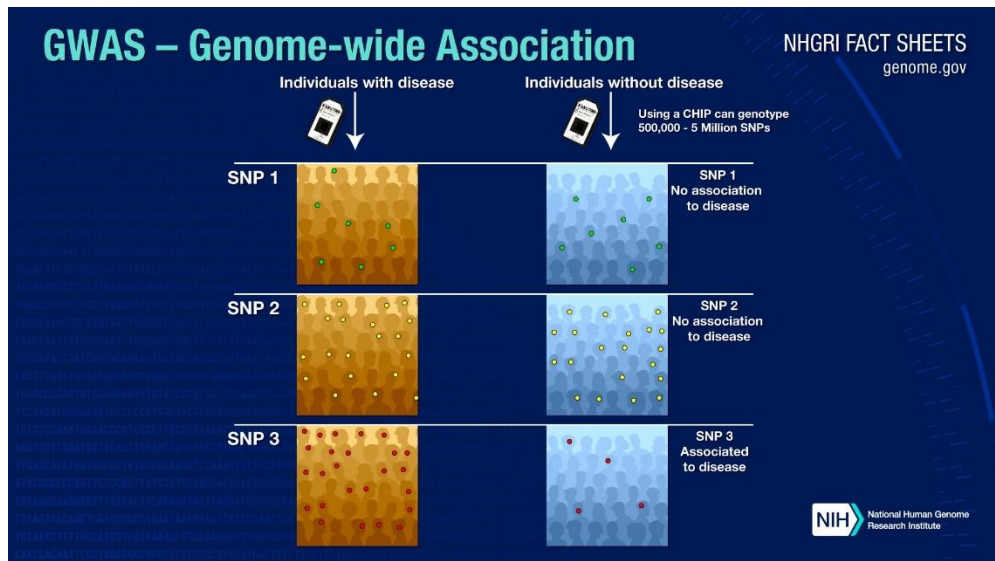
3. Why doesn't increased read quantity increase the quality of the data?

Article #3 Notes: Advantages of paired-end and single-read sequencing

Source Title	Advantages of paired-end and single-read sequencing
Source citation (APA Format)	<p>Illumina. (n.d.). <i>Paired-end vs. Single-read sequencing</i>. Illumina.com. Retrieved September 26, 2024, from https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html</p>
Original URL	https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html
Source type	Website Article
Keywords	Long-read, short-read, RNA-Seq, Genomics
#Tags	#lrRNAseq #srRNAseq #rna #genomics
Summary of key points + notes (include methodology)	<p>Illumina is a company that specializes in making machinery for RNA-Sequencing/NGS. This article reviews a key factor in every RNA-Seq study that needs to be considered before any experiment – the difference between paired-end and single-read sequencing. Single-read sequencing is when RNA is sequenced from only one end, simplifying the data output, and making it easier to deal with in terms of cost and usage. Paired-end sequencing works from both ends to gather higher quality data, among other observable characteristics such as gene fusions, novel transcripts, genomic rearrangements, and more. Overall, paired-end sequencing is more flexible, and it is more likely to return more significant or broader results.</p>
Research Question/Problem/Need	What are the differences of paired-end sequencing and single-read sequencing and which one should you pick?

<p>Important Figures</p>	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Paired-End Reads</p>  </div> <div style="text-align: center;"> <p>Alignment to the Reference Sequence</p>  </div> </div> <hr/> <p>Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.</p>
<p>VOCAB: (w/definition)</p>	<p>Alignable – In the context of genomics, the ability of RNA to match DNA cDNA – Complementary DNA, or DNA that is reverse transcribed</p>
<p>Cited references to follow up on</p>	<p>https://www.illumina.com/techniques/sequencing/rna-sequencing.html https://www.illumina.com/events/webinar/2021/ngs-cancer-research.html</p>
<p>Follow up Questions</p>	<ol style="list-style-type: none"> 1. How cost-effective is single-read compared to paired-end sequencing? 2. How likely are these methods to provide some sort of return or correlation? 3. What other resources does Illumina provide besides making the machines?

Article #4 Notes: GWAS Fact Sheet (NIH)

Source Title	Genome-Wide Association Studies Fact Sheet
Source citation (APA Format)	<i>Genome-wide association studies fact sheet.</i> (2019, March 9). Genome.gov; NHGRI. https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet
Original URL	https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet
Source type	Online Article
Keywords	GWAS, genetic associations, human genetic variation, genome
#Tags	#GWAS, #NIH, #factsheet, #NHGRI
Summary of key points + notes (include methodology)	This article from the National Human Genome Research Institute talks about genome-wide association studies (GWAS). In the context of healthcare, they are typically done to associate genetic markers with diseases, making them easier to identify and treat. The article then briefly describes different methods of conducting GWAS, ways to access publicly available data from GWAS, and the NIH's role in these types of experiments.
Research Question/Problem/Need	What is a GWAS (genome-wide association study) and why is it useful?
Important Figures	 <p>The infographic illustrates the process of GWAS. It compares two groups: 'Individuals with disease' (represented by orange circles) and 'Individuals without disease' (represented by blue circles). Three SNPs are analyzed: SNP 1, SNP 2, and SNP 3. For SNP 1 and SNP 2, the distribution of alleles is similar in both groups, indicating 'No association to disease'. For SNP 3, the 'Individuals with disease' group has a higher frequency of a specific allele (red dots) compared to the 'Individuals without disease' group, indicating 'SNP 3 Associated to disease'. A note states 'Using a CHIP can genotype 500,000 - 5 Million SNPs'. The NIH logo and 'National Human Genome Research Institute' are also present.</p>
VOCAB: (w/definition)	<p>GWAS</p> <ul style="list-style-type: none"> - Genome-wide association studies are a type of genetic study in which the

	<p>genome is scanned to find genetic variations associated with the disease.</p> <p>Genome</p> <ul style="list-style-type: none"> - Complete sets of DNA <p>Genetic marker</p> <ul style="list-style-type: none"> - A DNA sequence with a known physical location on a chromosome
Cited references to follow up on	<p>http://www.nhlbi.nih.gov/news/press-releases/2006/nhlbi-to-launch-framingham-genetic-research-study.html -- Long-term heart study involving GWAS technologies</p>
Follow up Questions	<ol style="list-style-type: none"> 1. What are specific ways we can take knowing about the genes associated with disease to a cure? 2. How can we use novel technologies like RNA-Seq along with GWAS data to find out more about these genetic markers? 3. Could GWAS be useful in categorizing diseases we may not yet know about? (ex. Varying subsections of the same disease varying the symptoms, new diseases out of those which have already been categorized under one phenotype, etc.)

Article #5 Notes: Optimality of Graphlet Screening in High Dimensional Variable Selection

Source Title	Optimality of Graphlet Screening in High Dimensional Variable Selection
Source citation (APA Format)	Jin, J., Zhang, C.-H., & Zhang, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. <i>Journal of Machine Learning Research: JMLR</i> , 15(79), 2723–2772. https://www.jmlr.org/papers/v15/jin14a.html
Original URL	https://www.jmlr.org/papers/v15/jin14a.html
Source type	Journal Article
Keywords	High Dimensional Variable Selection, Graphlet Screening, Screen and Clean, Kai ² , Statistical Algorithm, Linear Regression
#Tags	#statistics #algorithm #linearRegression #matlab #R
Summary of key points + notes (include methodology)	Variable selection at its most basic level involves finding non-zero coordinates to a system to streamline the formation of a prediction (regression) model by determining all relevant variables in a given situation. Graphlet Screening detects signals that are both rare (as in only a certain section of the data is non-zero) and weak (<i>lacking in correlation</i>) using a screen (kai ² tests of subgraphs) and clean (hamming distance as a loss function & penalized MLE) method. These signals are found commonly in GWAS and NGS studies, and most other existing methods don't work on these types of signals, favoring data that is sparse and <i>strong</i> instead. Therefore, a method such as Graphlet Screening is uniquely favorable for considering which factors influence genetic outcomes.
Research Question/Problem/Need	How can we form linear regression models based on data which is rare (lots of non-zero data) and individually weak (lacking strong correlation)?
Important Figures	None
VOCAB: (w/definition)	Graphlet Screening – Screen and clean method to tackle high dimensional variable selection. Hamming Distance – A way to calculate distance between integers/binary string, used as a loss function Sparsity – How much of the data contains non-zero values. A “sparse” dataset is generally agreed to be a data set which contains mostly non-zero values. Correlation – Relation between variables and model
Cited references to follow up on	None
Follow up Questions	- What does it mean when the signals/selected variables are “individually weak” in the context of the paper?

- | | |
|--|--|
| | <ul style="list-style-type: none">- Does this advocate for linear dependent variables as well as linear independent variables?- Why can't other methods of selection be implemented for "sparse and weak" data? |
|--|--|

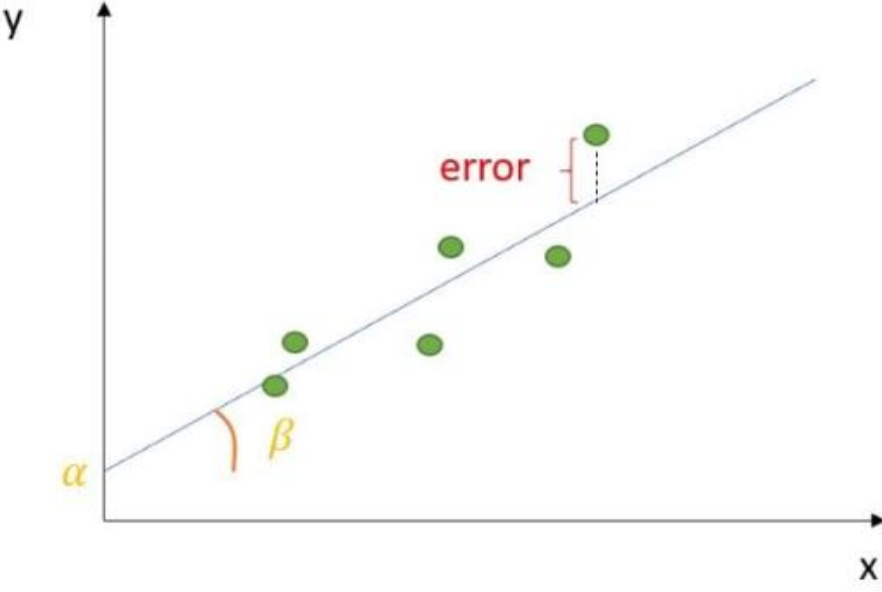
Article #6 Notes: What is a Linear Regression Model?

Source Title	What is a Linear Regression Model?
Source citation (APA Format)	Mathworks, & Simulink. (n.d.). <i>What is a Linear Regression model?</i> Mathworks.com. Retrieved October 3, 2024, from https://www.mathworks.com/help/stats/what-is-linear-regression.html
Original URL	https://www.mathworks.com/help/stats/what-is-linear-regression.html
Source type	Website Article
Keywords	Statistics, Linear regression, Statistical model
#Tags	#stats #regression #statsmodel
Summary of key points + notes (include methodology)	<p>A linear regression model demonstrates a quantifiable relationship between a dependent variable/matrix Y and (usually) many independent variables in the form of a matrix X, also known as the design matrix.</p> <p>This usually involves calculating an underlying relationship between the independent variable and the dependent variable (Beta) and adding noise to account for random error (as beta can never be completely accurate). This relationship between $\beta \cdot x$ and y is linear, hence the name “linear” regression.</p>
Research Question/Problem/Need	What is a Linear Regression model?
Important Figures	<p>A multiple linear regression model is</p> $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \dots, n,$ <p>where</p> <ul style="list-style-type: none"> • n is the number of observations. • y_i is the ith response. • β_k is the kth coefficient, where β_0 is the constant term in the model. your design matrix X. • X_{ij} is the ith observation on the jth predictor variable, $j = 1, \dots, p$. • ε_i is the ith noise term, that is, random error. <p>Figure 1. Mathematical model representing in general the format of a linear regression model.</p>
VOCAB: (w/definition)	<p>Design matrix – matrix containing independent variables to be assessed for association with Y matrix (X matrix)</p> <p>Regression – the general relationship between two variables</p>

Cited references to follow up on	https://www.mathworks.com/help/stats/linear-regression-model-workflow.html https://www.mathworks.com/help/stats/linearmodel.html
Follow up Questions	<ol style="list-style-type: none">1. What are the different methods of obtaining the estimate for beta? How are some better than others?2. How do methods differ in determining associated variables work based on the analysis of unique data sets?3. How do we determine how far our estimate for beta is without beta itself?

Article #7: Understanding Ordinary Least Squares (OLS) Regression

Source Title	Understanding Ordinary Least Squares (OLS) Regression
Source citation (APA Format)	Alto, V. (2023, February 14). Understanding Ordinary Least Squares (OLS) Regression. <i>Built In</i> . https://builtin.com/data-science/ols-regression
Original URL	https://builtin.com/data-science/ols-regression
Source type	Newspaper Article
Keywords	Statistics, Regression model, Linear regression, Statistical model, Ordinary Least Squares
#Tags	#stats #regression #ols #statsmodel
Summary of key points + notes (include methodology)	<p>Ordinary Least Squares (OLS) is a method of predicting beta in a multiple linear regression model (underlying relationship between many variables to one response variable). It does this mainly by minimizing the square error for beta and alpha (y-intercept of the linear function).</p> <p>OLS is a good estimation method for data sets which have stronger, unbiased correlations as it can output unbiased estimates for alpha and beta. However, it may lack performance in data sets which do not have as strong correlations and in datasets which have too many variables.</p>
Research Question/Problem/Need	What methods can we use to estimate Beta in linear regression?

Important Figures	 <p>Figure 1. This graph represents the role of alpha (the y-intercept) and b(the slope as an angle) for a single variable as a part of linear regression.</p>
VOCAB: (w/definition)	<p>OLS – Ordinary Least Squares Classification – Splitting up data or making predictions categorically. Multiple linear regression – Linear regression across many different design variables Design variables – values within the design matrix</p>
Cited references to follow up on	<p>None</p>
Follow up Questions	<ol style="list-style-type: none"> 1. What other algorithms exist? 2. Why does the method break at higher dimensional variables? 3. What is the significance of alpha? Why do some models not consider it?

Article #8: Genomic regression analysis of coordinated expression

Source Title	Genomic regression analysis of coordinated expression
Source citation (APA Format)	Cai, L., Li, Q., Du, Y., Yun, J., Xie, Y., DeBerardinis, R. J., & Xiao, G. (2017). Genomic regression analysis of coordinated expression. <i>Nature Communications</i> , 8(1). https://doi.org/10.1038/s41467-017-02181-0
Original URL	https://www.nature.com/articles/s41467-017-02181-0
Source type	Journal Article
Keywords	Genomics, Statistics, Genomic Analysis, Regression
#Tags	#genomics #statistics #linearRegression #rnaSeq
Summary of key points + notes (include methodology)	<p>Co-expression analysis (detection of co-expression signatures released by associated genes) is used to predict gene function and functionally related gene sets, but SCNAs (somatic copy number alterations) may interfere with this process specifically surrounding the study of cancer cells due to releasing similar signals.</p> <p>This study defines a method to adjust for SCNAs called GRACE (Genomic regression analysis of coordination expression). This can help tackle gene expression data gathered by RNA-seq experiments and associate the data with gene expression. It then tests this method using cancer cells which produce large amounts of SCNAs that are hard to measure using traditional methods.</p> <p>This is done by using a linear regression model with the copy number alteration values and the RNA levels as a response vector.</p>
Research Question/Problem/Need	How can we use linear regression to analyze co-expression?

Important Figures

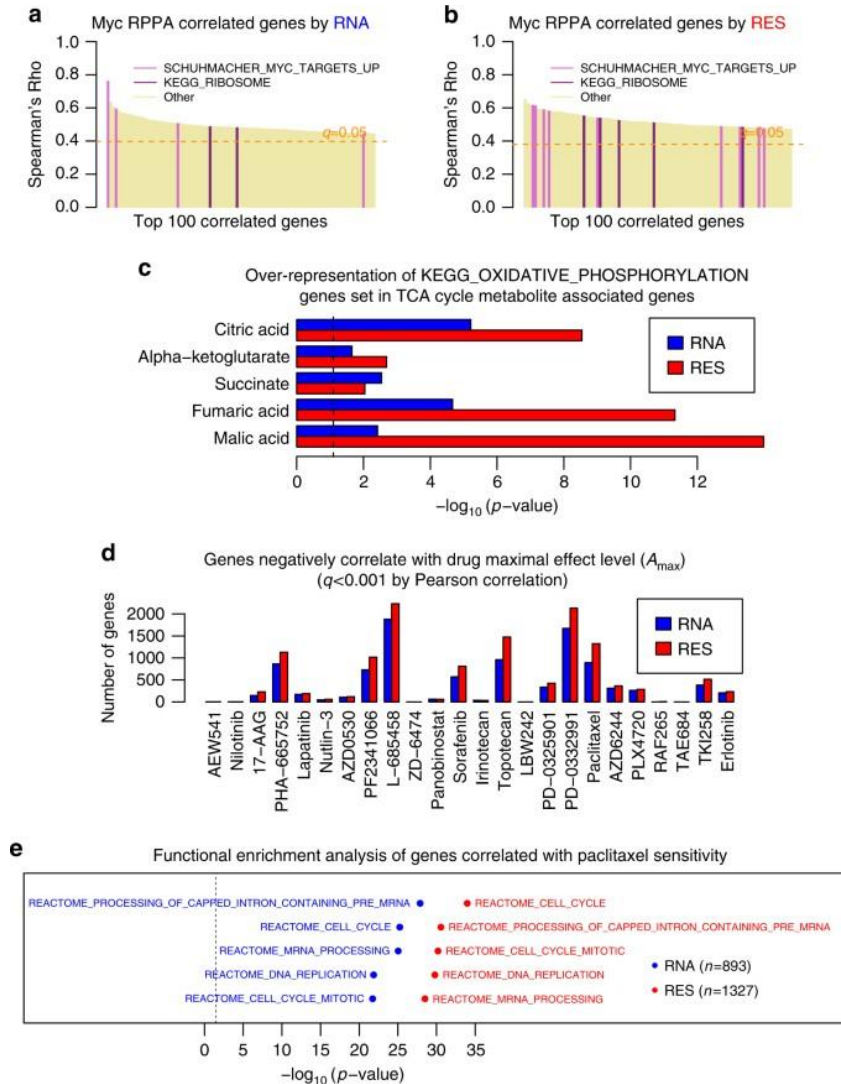


Figure 1. This figure observes the difference in detected correlated genes when considering perceived RNA levels and then considering the (RES)idual RNA levels. It considers which methods are overestimating the found correlations, and at which points the analysis provided negative correlations.

$$Y_j = \beta_j X_j + \alpha_{jh} Y_h + \epsilon_j$$

$$Y_h = \beta_h X_h + \alpha_{hj} Y_j + \epsilon_h,$$

Figure 2. This displays the linear regression models that were used to create the models which did the analysis for this study. Gaussian noise is added to a correlation coefficient alpha along with a residual method coefficient beta.

VOCAB: (w/definition)

SCNA – somatic copy number alterations, an evasive trait of cancer cells which affects the gene expression across many different functions by affecting their dosage.
 Co-expression – Correlations in transcript levels.

	Response vector – What a linear regression model aims to estimate or predict or “fit” to.
Cited references to follow up on	https://doi.org/10.1073%2Fpnas.162471999 https://doi.org/10.1371%2Fjournal.pone.0053014
Follow up Questions	<ol style="list-style-type: none">1. Why does the specific regression model used work so well in this context?2. Are there any other models that work well for distinguishing SCNAs from real RNA signals?3. Are there more benefits from RNA-Seq that can be carried over now that it can be properly analyzed with cancer cells?

Article #9: Variable selection and estimation in high-dimensional models

Source Title	Variable selection and estimation in high-dimensional models
Source citation (APA Format)	Horowitz, J.L. (2015), Variable selection and estimation in high-dimensional models. <i>Canadian Journal of Economics/Revue canadienne d'économique</i> , 48: 389-407. https://doi.org/10.1111/caje.12130
Original URL	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3092303/
Source type	Journal Article
Keywords	Linear Regression, Statistical models
#Tags	#linearRegression #statistics
Summary of key points + notes (include methodology)	<p>Variable selection is a crucial part of analyzing any data, as it has to do with associated variables and determining their correlation to certain outcomes. Identifying these correlations when the correlations are sparse (mostly non-zero) is estimated through many different models. This paper goes over some different methods of quantifying the correlations between a given design matrix and a response matrix and rates their performance through Monte Carlo simulations and empirical examples.</p> <p>Particularly, we want an algorithm that can deal with when the number of correlated variables exceeds the total sample size ($p > n$). An example of an algorithm that doesn't work for this is Ordinary Least Squares.</p> <p>The usage of LASSO, adaptive LASSO, the bridge penalty function, SCAD (smoothly clipped absolute deviation) penalty, the minmax concave penalty function, linear-in-parameters quantile regression model, and other penalized modeling systems are reviewed through the paper.</p>
Research Question/Problem/Need	How do we use linear regression models when the correlated variables exceeds the sample size?

Important Figures	<p>Table 1. Results of Monte Carlo experiments</p> <table border="1" data-bbox="535 268 1487 548"> <thead> <tr> <th>d</th> <th>MSE OLS</th> <th>MSE OLS with true model</th> <th>MSE LASSO</th> <th>MSE AL</th> <th>LASSO SIZE</th> <th>AL SIZE</th> <th>LASSO PROB LARGE</th> <th>AL PROB LARGE</th> </tr> </thead> <tbody> <tr> <td>2</td> <td>0.67</td> <td>0.22</td> <td>0.27</td> <td>0.19</td> <td>7.9</td> <td>5.8</td> <td>0.88</td> <td>0.67</td> </tr> <tr> <td>4</td> <td>0.67</td> <td>0.19</td> <td>0.29</td> <td>0.17</td> <td>10.6</td> <td>8.0</td> <td>0.81</td> <td>0.64</td> </tr> <tr> <td>6</td> <td>0.67</td> <td>0.16</td> <td>0.40</td> <td>0.19</td> <td>13.3</td> <td>10.2</td> <td>0.67</td> <td>0.43</td> </tr> </tbody> </table> <p>Figure 1. This table explores the results of the study. The columns labelled with MSE (mean squared errors) represent how far off the correlation estimations were in comparison to the actual correlations, so smaller is better here. AL (adaptive LASSO) generally has the lowest MSE out of the other model options (LASSO, OLS with true model, OLS). Columns 6 and 7 show the average number of covariates for both AL and LASSO, while columns 8 and 9 shows the probability that these methods contain all of the covariates with large coefficients.</p>	d	MSE OLS	MSE OLS with true model	MSE LASSO	MSE AL	LASSO SIZE	AL SIZE	LASSO PROB LARGE	AL PROB LARGE	2	0.67	0.22	0.27	0.19	7.9	5.8	0.88	0.67	4	0.67	0.19	0.29	0.17	10.6	8.0	0.81	0.64	6	0.67	0.16	0.40	0.19	13.3	10.2	0.67	0.43
d	MSE OLS	MSE OLS with true model	MSE LASSO	MSE AL	LASSO SIZE	AL SIZE	LASSO PROB LARGE	AL PROB LARGE																													
2	0.67	0.22	0.27	0.19	7.9	5.8	0.88	0.67																													
4	0.67	0.19	0.29	0.17	10.6	8.0	0.81	0.64																													
6	0.67	0.16	0.40	0.19	13.3	10.2	0.67	0.43																													
VOCAB: (w/definition)	<p>Sparse – mostly non-zero Monte Carlo simulation – statistical technique of predicting the possible outcomes given a certain event</p>																																				
Cited references to follow up on	<p>https://www.tandfonline.com/doi/abs/10.1198/016214501753382273 https://www.jstor.org/stable/24310359</p>																																				
Follow up Questions	<ol style="list-style-type: none"> 1. Which one of these models is most compatible with genomic data? 2. How do these methods compare to Graphlet Screening? 3. When specifically should I use non-linear methods of regression? 																																				

Article #10: Efficient Regularized Regression with L_0 Penalty for Variable Selection and Network Construction

Source Title	Efficient Regularized Regression with L_0 Penalty for Variable Selection and Network Construction																																									
Source citation (APA Format)	Liu, Z., & Li, G. (2016). Efficient regularized regression with L_0 Penalty for variable selection and network construction. <i>Computational and Mathematical Methods in Medicine</i> , 2016, 1–11. https://doi.org/10.1155/2016/3456153																																									
Original URL	https://onlinelibrary.wiley.com/doi/full/10.1155/2016/3456153																																									
Source type	Journal Article																																									
Keywords	Linear regression, Statistical																																									
Tags	#linearRegression #statistics #statsmodel																																									
Summary of key points + notes (include methodology)	<p>L_0-Penalized Maximum Likelihood estimation is a method of determining the parameters or correlated variables (variable selection) for a given linear regression model. Although it is effective, it is computationally challenging and not worth it in comparison to other variable selection methods. This paper outlines two methods of estimating the L_0-Penalization method, improving the efficiency such as lasso, SCAD, and MC+.</p> <p>The L_0-Penalization is used in conjunction with the Expectation-Maximization algorithm (EM) to create a L_0EM model and a dual L_0EM (DL_0EM) that is more computationally efficient.</p>																																									
Research Question/Problem/Need	How do we make existing methods of linear regression more efficient for use in larger datasets?																																									
Important Figures	<table border="1"> <thead> <tr> <th rowspan="2">r</th> <th colspan="3">L_0</th> <th colspan="3">L_1</th> </tr> <tr> <th># SF</th> <th>MSE</th> <th>$\ \hat{\theta} - \theta\$</th> <th># SF</th> <th>MSE</th> <th>$\ \hat{\theta} - \theta\$</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>3.39 (± 1.1)</td> <td>1.01 (± 0.14)</td> <td>0.206 (± 0.12)</td> <td>14.5 (± 3.45)</td> <td>1.19 (± 0.19)</td> <td>0.38 (± 0.1)</td> </tr> <tr> <td>0.3</td> <td>3.37 (± 0.9)</td> <td>1.02 (± 0.16)</td> <td>0.23 (± 0.12)</td> <td>14.5 (± 2.91)</td> <td>1.21 (± 0.19)</td> <td>0.41 (± 0.19)</td> </tr> <tr> <td>0.6</td> <td>3.49 (± 1.7)</td> <td>1.02 (± 0.23)</td> <td>0.23 (± 0.16)</td> <td>13.5 (± 3.0)</td> <td>1.26 (± 0.2)</td> <td>0.54 (± 0.15)</td> </tr> <tr> <td>0.8</td> <td>3.32 (± 0.9)</td> <td>1.06 (± 0.15)</td> <td>0.28 (± 0.21)</td> <td>11.7 (± 2.69)</td> <td>1.3 (± 0.21)</td> <td>0.89 (± 0.25)</td> </tr> </tbody> </table> <p>Figure 1. This table displays the difference between two regression methods using multiple simulations. One is derived from the modified L_0 estimation and the other is the L_1-penalized estimation method. # SF refers to the average number of</p>	r	L_0			L_1			# SF	MSE	$\ \hat{\theta} - \theta\ $	# SF	MSE	$\ \hat{\theta} - \theta\ $	0	3.39 (± 1.1)	1.01 (± 0.14)	0.206 (± 0.12)	14.5 (± 3.45)	1.19 (± 0.19)	0.38 (± 0.1)	0.3	3.37 (± 0.9)	1.02 (± 0.16)	0.23 (± 0.12)	14.5 (± 2.91)	1.21 (± 0.19)	0.41 (± 0.19)	0.6	3.49 (± 1.7)	1.02 (± 0.23)	0.23 (± 0.16)	13.5 (± 3.0)	1.26 (± 0.2)	0.54 (± 0.15)	0.8	3.32 (± 0.9)	1.06 (± 0.15)	0.28 (± 0.21)	11.7 (± 2.69)	1.3 (± 0.21)	0.89 (± 0.25)
r	L_0			L_1																																						
	# SF	MSE	$\ \hat{\theta} - \theta\ $	# SF	MSE	$\ \hat{\theta} - \theta\ $																																				
0	3.39 (± 1.1)	1.01 (± 0.14)	0.206 (± 0.12)	14.5 (± 3.45)	1.19 (± 0.19)	0.38 (± 0.1)																																				
0.3	3.37 (± 0.9)	1.02 (± 0.16)	0.23 (± 0.12)	14.5 (± 2.91)	1.21 (± 0.19)	0.41 (± 0.19)																																				
0.6	3.49 (± 1.7)	1.02 (± 0.23)	0.23 (± 0.16)	13.5 (± 3.0)	1.26 (± 0.2)	0.54 (± 0.15)																																				
0.8	3.32 (± 0.9)	1.06 (± 0.15)	0.28 (± 0.21)	11.7 (± 2.69)	1.3 (± 0.21)	0.89 (± 0.25)																																				

	selected features while MSE refers to the mean squared error. The fourth and seventh column are referring to the average bias for each method.
VOCAB: (w/definition)	NP-hard – Refers to when the program time efficiency is polynomial/computationally unviable. MC+: Minmax concave, a regression method.
Cited references to follow up on	https://academic.oup.com/jrsssb/article/58/1/267/7027929 https://www.tandfonline.com/doi/abs/10.1198/106186007X255676
Follow up Questions	<ol style="list-style-type: none"> 1. Is L_0 used in a genomic context? 2. How are linear regression models generally modified to fit the field of context? 3. Can this process RNA-Seq data?

Patent #1: Artificial Intelligence-Based Sequencing

Source Title	Artificial Intelligence-Based Sequencing
Source citation (APA Format)	Jaganathan, K., Dutta, A., Kashefhighi, D., Gobbel, J. R., & Amirali, K. I. A. (2020). <i>Artificial intelligence-based sequencing</i> (Patent No. 20200302224:A1). In US Patent (20200302224:A1).
Original URL	https://patents.google.com/patent/US20100159533A1/en
Source type	Patent
Keywords	Patent, Sequencing, Genomics, Linear Regression
#Tags	#patent #rnaSeq #sequencing #genomics #linearRegression
Summary of key points + notes (include methodology)	<p>This patent was filed and is active under Illumina Inc. They create machines to do RNA-Sequencing and other genomics related tasks. This patent specifically uses neural networks and linear regression to digitally implement end-to-end sequencing, or paired-end sequencing.</p> <p>In the first layer of the two neural networks, an image captured by a sequencing system that contains clusters produced through BFS (Breadth-first-search) is passed into a neural network to generate a cluster map, identifying traits of the clusters such as size, shapes, etc. The second layer involves a second image with slightly modified intensity levels based on the first image, which is again passed to a neural network. By checking the light intensity in certain areas, the neural networks can sequence the collected data from the initial clusters.</p>
Research Question/Problem/Need	How can we take images obtained by high-precision sequencing machinery and actually sequence them according to their clusters?

Important Figures

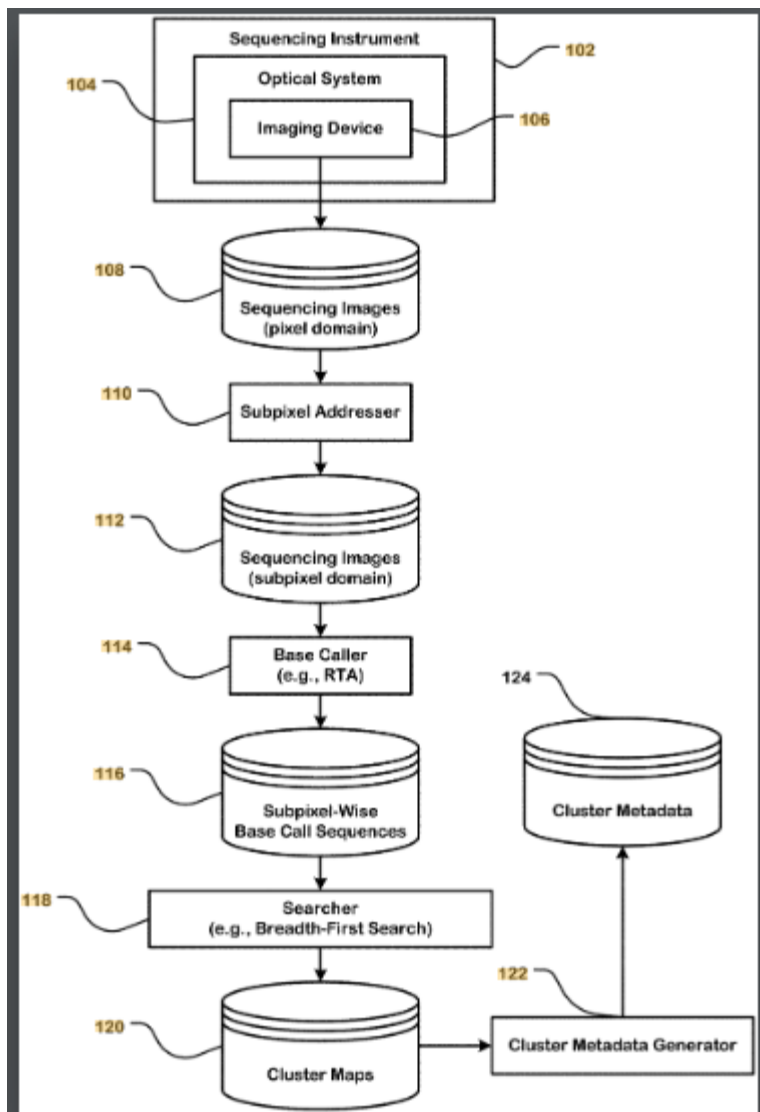


Figure 1. This flowchart displays the process described by the patent in which the images collected by the sequencing instrument are used to identify clusters which can then be used to be create cluster maps which can be used to aid in Sequencing.

VOCAB: (w/definition)	BFS – Breadth-first search, searches a data tree for nodes which satisfy a given property, important part of dividing subgraphs to identify correlations.
Cited references to follow up on	https://patents.google.com/patent/WO2023220627A1/en
Follow up Questions	<ol style="list-style-type: none"> 1. Can neural networks also be used in variable selection? 2. Should neural networks be considered when tacking genomic data? 3. Why are multiple models necessary?

Patent #2: Deep learning-based techniques for pre-training deep convolutional neural networks

Source Title	Deep learning-based techniques for pre-training deep convolutional neural networks
Source citation (APA Format)	Farh, K.-H., Gao, H., & Padigepati, S. R. (2021). <i>Deep learning-based techniques for pre-training deep convolutional neural networks</i> (Patent No. AU:2019272062:B2). In <i>Patent</i> (AU:2019272062:B2).
Original URL	https://patents.google.com/patent/AU2019272062B2/en
Source type	Patent
Keywords	Patent, Deep learning, Genomics, Linear Regression, Statistics
#Tags	#statistics #genomics #linearRegression #deepLearning
Summary of key points + notes (include methodology)	Neural networks are used often to process sequences of genomic data and amino acids during analysis. However, they can sometimes overfit the data (adhere too closely to training data and be unable to make predictions). To fix this problem, this patent uses position frequency matrices (PFMs) as supplemental training data to influence the model's perception of correlation and keep the model from overfitting to the training data (which is assumed to be genomic sequencing data). There is also coded logic to apply the training PFMs to the neural network and precisely determine what the influence of the PFMs should be.
Research Question/Problem/Need	How do we prevent overfitting when analyzing genetic data?
Important Figures	None
VOCAB: (w/definition)	PFMs – Position frequency matrix, represents the frequency of each nucleotide or amino acid within a sequence alignment.
Cited references to follow up on	https://patents.google.com/patent/CN108197427B/en
Follow up Questions	<ol style="list-style-type: none"> 1. How can this type of neural network be used in contexts outside of genomics? 2. What are the direct causes of overfitting? 3. Are there other methods to prevent overfitting that are more efficient?

Article #11: LASSO regression

Source Title	LASSO regression
Source citation (APA Format)	Ranstam, J., & Cook, J. A. (2018). LASSO regression. <i>The British Journal of Surgery</i> , 105(10), 1348–1348. https://doi.org/10.1002/bjs.10895
Original URL	https://academic.oup.com/bjs/article/105/10/1348/6122951
Source type	Journal Article
Keywords	Linear regression, variable selection
#Tags	#regression #stats #algorithm #linearRegression
Summary of key points + notes (include methodology)	<p>LASSO has shown, at least in the past, that it was at some point very effective at genomic analysis. It works well by tackling overfitting in high-dimensional variable selection scenarios.</p> <p>Although it performs generally well in some scenarios, it still has problems with overfitting and optimism bias. It also may introduce bias on individual variables or characteristics to achieve a better overall result. This, by extension, makes it difficult to use with interactive factors in the genome.</p>
Research Question/Problem/Need	How do we analyze the genome's correlation to phenotypes using regression? What is industry standard and what are its benefits and its vices?
Important Figures	None.
VOCAB: (w/definition)	<p>Interactive factors – when there are multiple pieces of DNA that have to be mutated in a specific orientation in order to reliably impact the phenotype</p> <p>Overfitting – when the regression model fits too closely to the training data and loses its ability to make reliable predictions</p> <p>Optimism bias – when a signal is really weak, and could be 0, but there is a small amount of evidence that it could be valid, the regression will include the factor even though there is a high likelihood that it will result in error.</p>
Cited references to follow up on	https://academic.oup.com/jrsssb/article/67/2/301/7109482
Follow up Questions	<ol style="list-style-type: none"> 1. How is error in LASSO typically measured? 2. Is it okay to sacrifice individual correlations for overall better results? 3. Why do researchers use LASSO over L0?

Article #12: Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods

Source Title	Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods
Source citation (APA Format)	Qin, J., Hu, Y., Xu, F., Yalamanchili, H. K., & Wang, J. (2014). Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. <i>Methods (San Diego, Calif.)</i> , 67(3), 294–303. https://doi.org/10.1016/j.ymeth.2014.03.006
Original URL	https://www.sciencedirect.com/science/article/pii/S1046202314000991?via%3Dihub
Source type	Journal Article
Keywords	Linear regression, variable selection
#Tags	#regression #stats #algorithm #linearRegression
Summary of key points + notes (include methodology)	<p>Gene expression is hard to quantify directly from the genome, which impacts phenotypes in complex and interactive ways. LASSO regression is commonly used within the field of genomic analysis to quantify these connections.</p> <p>However, LASSO is not accurate for larger genomes. This study uses an L0 algorithm and compares it to LASSO and finds that it is far more effective than LASSO when determining the gene expression data of mouse embryonic stem cells, obtained from a public, pre-collected database.</p> <p>It utilizes Chromatin immunoprecipitation (ChIP), a technique that studies the interaction between DNA and proteins. By sequencing this data through this method, all 3 algorithms perform better. Especially when integrating ChIP-X data, the algorithms performed significantly better than the LASSO. However, when this data is not integrated, the results are only slightly more accurate.</p>
Research Question/Problem/Need	Is L0 more capable than LASSO as an analysis method?

Important Figures

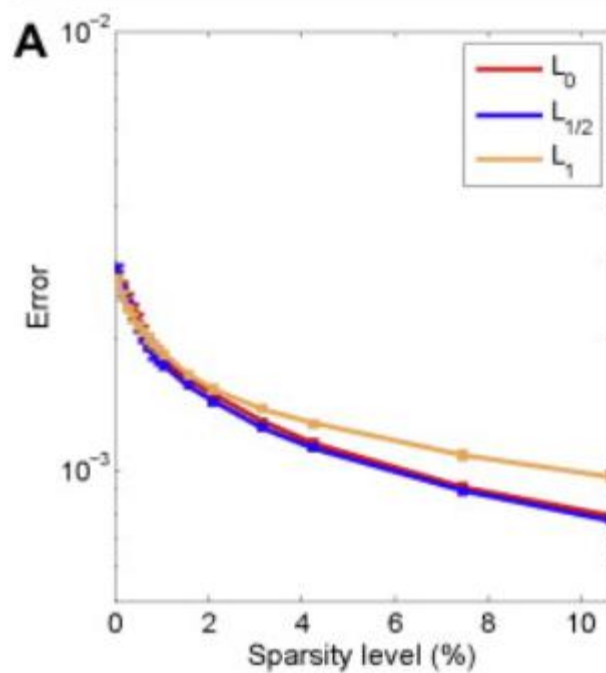


Figure 1 contains a graph displaying how as the sparsity % increases (signals increase), the error tends to decrease among all tested variations of L_0 . $L_{1/2}$ leads very slightly ahead of L_0 , which is far ahead of L_1 (LASSO).

VOCAB: (w/definition)

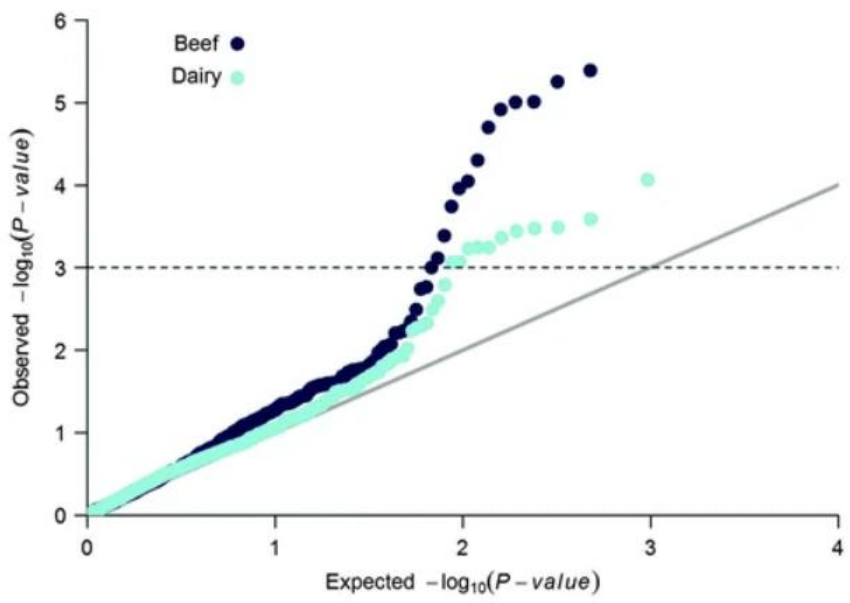
Chromatin immunoprecipitation – A technique of linking the genome to protein production

Cited references to follow up on

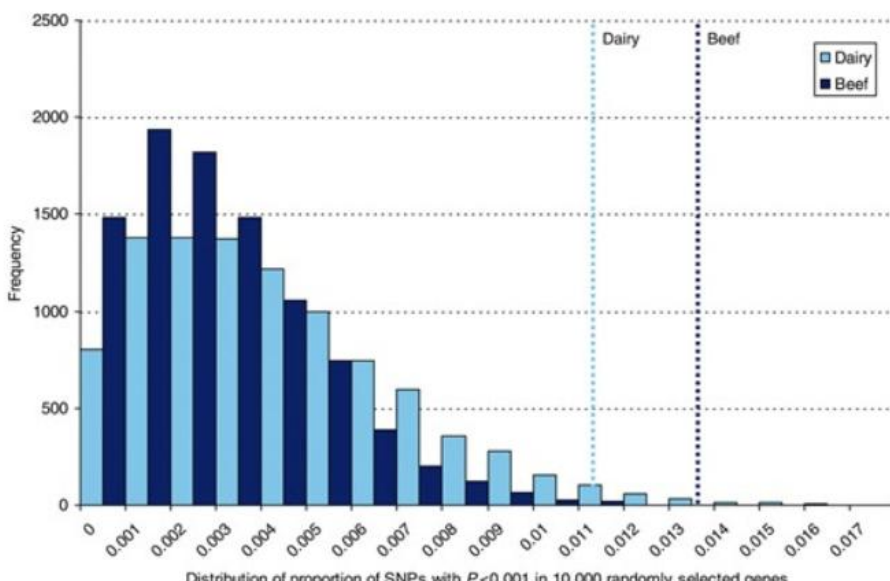
<https://doi.org/10.1093/nar/gkr593>

Article #13: Overview of Statistical Methods for Genome-Wide Association Studies (GWAS)

Source Title	Overview of Statistical Methods for Genome-Wide Association Studies (GWAS)
Source citation (APA Format)	Hayes, B. (2013). Overview of statistical methods for genome-wide association studies (GWAS). In <i>Methods in Molecular Biology</i> (pp. 149–169). Humana Press. doi.org/10.1007/978-1-62703-447-0_6
Original URL	https://link.springer.com/protocol/10.1007/978-1-62703-447-0_6
Source type	Protocol
Keywords	GWAS, Population Structure, Multiple Testing
#Tags	
Summary of key points + notes (include methodology)	<p>This article is an overview on how to do a GWAS, particularly the statistical methods necessary for doing a GWAS. It also discusses the method of obtaining the appropriate significance level for GWAS testing.</p> <p>GWAS analysis mainly contains any variation of linear regression which links the design matrix to the phenotype (we know this already). This article discusses both single marker regression (where a single genetic variation is compared) and haplotype analysis (where many genetic variations under the same chromosome are compared, which we are interested in).</p> <p>In situations with multiple genetic types and even in single marker regression, determining the right p-value (significance level) is crucial to</p>
Research Question/Problem/Need	What different methods of analysis are used when analyzing GWAS (genome-wide association studies)?

Important Figures	 <p>Figure 1. This figure describes a quartile-quartile plot for a certain GWAS study (Pryce et al.) in which the stature (height) of cattle is measured according to the “dairy” and “beef” populations. The idea is that since the P value exceeds the expected p value line (the positive slope full line), there is a correlation between the population and the phenotype. The horizontal dashed line represents the significance level of $\alpha = 0.001$.</p>
VOCAB: (w/definition)	None
Cited references to follow up on	https://doi.org/10.1534%2Fgenetics.110.123943
Follow up Questions	<ol style="list-style-type: none"> 1. The Graphlet Screening method tends to be more helpful for high-dimensional analysis, in which there are multiple interactive genomic factors that influence the phenotype. What advantage(s) do these multiple regression studies have over single marker regression studies? 2. How would I calculate the appropriate significance level for my experiment? 3. How many different variants of GWAS exist generally?

Article #14: Polymorphic Regions Affecting Human Height Also Control Stature in Cattle

Source Title	Polymorphic Regions Affecting Human Height Also Control Stature in Cattle
Source citation (APA Format)	Pryce, J. E., Hayes, B. J., Bolormaa, S., & Goddard, M. E. (2011). Polymorphic regions affecting human height also control stature in cattle. <i>Genetics</i> , 187(3), 981–984. https://doi.org/10.1534/genetics.110.123943
Original URL	https://academic.oup.com/genetics/article/187/3/981/6063280
Source type	Journal Article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	This study utilizes a GWAS to associate traits from cattle that have to do with height (stature). It found that the genes associated with both dairy and beef cattle were remarkably similar to the genes associated within humans. To analyze the data, associate phenotypes and retrieve a p-value, each SNP was regressed using their own custom model which mixed pedigrees with variations and correlations.
Research Question/Problem/Need	What genes affect stature in cattle, and are they the same as humans?
Important Figures	 <p>Figure 1. This figure shows a frequency chart which compares the distributions of significant ($p < 0.001$) SNPs detected for association with stature out of all SNPs</p>

	tested. Essentially, this means that the SNPs identified by the study are generally accurate and significant for cattle stature.
VOCAB: (w/definition)	None
Cited references to follow up on	https://www.publish.csiro.au/an/EA08273
Follow up Questions	<ol style="list-style-type: none">1. Do most genes between cattle and humans match?2. How are they determining the p-value?3. How does the fisher method compare?

Article #15: Variable selection in linear regression models: Choosing the best subset is not always the best choice

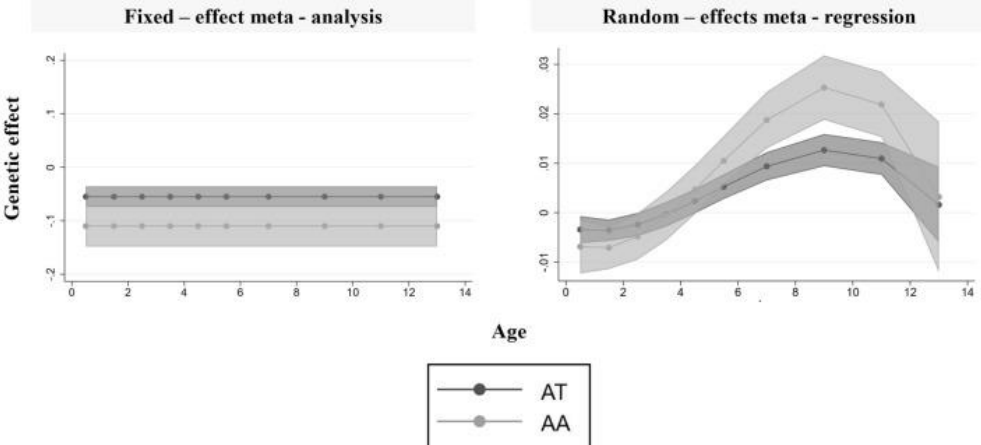
Source Title	Variable selection in linear regression models: Choosing the best subset is not always the best choice
Source citation (APA Format)	Hanke, M., Dijkstra, L., Foraita, R., & Didelez, V. (2024). Variable selection in linear regression models: Choosing the best subset is not always the best choice. <i>Biometrical Journal</i> , 66(1). https://doi.org/10.1002/bimj.202200209
Original URL	https://onlinelibrary.wiley.com/doi/full/10.1002/bimj.202200209
Source type	Journal Article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	<p>This article discusses various methods of high dimensional variable selection (aka identifying correlated features between a design matrix and a response vector). Particularly, it addresses a method called “best subset selection” (BSS) which, until recently, was an NP-Hard algorithm (aka, computationally inefficient).</p> <p>Some existing proposals have made some (minimally tested) models that represent BSS as a non-NP-Hard algorithm, reinitializing its potential use case. As such, this study compares BSS to forward stepwise selection (FSS), LASSO, and Elastic net (Enet).</p> <p>The study, in short, found that BSS wasn’t actually as effective as these other methods. In fact, FSS performed fairly similar to the BSS model. The only advantage that the model had were in datasets with weak correlation and high levels of noise, even when the dimensionality of the data was low (which is what the model is claimed to be good at).</p>
Research Question/Problem/Need	Is the BSS model viable as an alternative now that it is more computationally efficient?
Important Figures	None
VOCAB: (w/definition)	None
Cited references to follow up on	https://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15
Follow up Questions	<ol style="list-style-type: none"> 1. How is Enet different from LASSO? 2. How have these models optimized BSS to become not NP-hard? Could

these methods be applied to things like LO?

3. Are there other methods like BSS that are waiting to be optimized? What is the demand for this type of research?

Article #16: Meta-regression of genome-wide association studies to estimate age-varying genetic effects

Source Title	Meta-regression of genome-wide association studies to estimate age-varying genetic effects
Source citation (APA Format)	Pagoni, P., Higgins, J. P. T., Lawlor, D. A., Stergiakouli, E., Warrington, N. M., Morris, T. T., & Tilling, K. (2024). Meta-regression of genome-wide association studies to estimate age-varying genetic effects. <i>European journal of epidemiology</i> , 39(3), 257–270. https://doi.org/10.1007/s10654-023-01086-1
Original URL	https://pmc.ncbi.nlm.nih.gov/articles/PMC10995067/
Source type	Journal Article
Keywords	Genome-wide association studies, Meta-analysis, Meta-regression, Age-varying genetic effects
#Tags	
Summary of key points + notes (include methodology)	<p>This study does a meta-regression of various GWAS to estimate the changes in genetic effects as people age. It does this to link interactions between age and SNPs.</p> <p>Particularly, it found that when applied to BMI, the influence of certain SNPs changed as people aged. Generally, this proves meta-regression-analysis as a helpful method for analyzing GWAS results and connecting them to other studies as well. Additionally, this also clarifies that genetic effects aren't actually consistent across all ages, and more GWAS needs to be done.</p> <p>This study compares two methods of meta-analysis: Fixed-Effect Meta-Analysis and Random-Effects Meta Regression. Fixed Effect assumes that the true effect size (amount of correlation) is the same across all studies and that any variations are because of collection error, weighting studies by their sample sizes and variability. Meanwhile, Random-Effects accounts for differences between studies by assuming that this correlation varies across studies and uses other factors to model how other factors may be involved (in the case of this study, age).</p>
Research Question/Problem/Need	How effective is meta-regression-analysis for combining GWAS results and making overall result comparisons?

Important Figures	<p>Estimated genetic association between rs9939609 SNP at the <i>FTO</i> locus and BMI</p>  <p>Figure 1. This figure compares two methods of meta-analysis, fixed-effect and random-effects meta regression. In the Random-Effects model, age is taken as a factor between studies and observes that peak correlation/effect from the rs9939609 SNP is at around the 9-year-old age mark. In this case, it shows that carriers of minor alleles showed lower BMI during infancy and higher BMI during childhood/adolescence, a factor that Fixed-Effect analysis does not factor in.</p>
VOCAB: (w/definition)	None.
Cited references to follow up on	https://www.nature.com/articles/nrg3472
Follow up Questions	<ol style="list-style-type: none"> 1. What other meta-analysis methods exist? 2. Can the Random-Effects model use multiple external parameters? 3. Can studies be reliably correlated into meta-analyses, or is this currently too unreliable for general practice/acceptance?

Article #17: High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in Genome Wide Association Studies

Source Title	High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in Genome Wide Association Studies
Source citation (APA Format)	Goudey, B., Abedini, M., Hopper, J. L., Inouye, M., Makalic, E., Schmidt, D. F., Wagner, J., Zhou, Z., Zobel, J., & Reumann, M. (2015). High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in Genome Wide Association Studies. <i>Health information science and systems</i> , 3(Suppl 1 HISA Big Data in Biomedicine and Healthcare 2013 Con), S3. https://doi.org/10.1186/2047-2501-3-S1-S3
Original URL	https://pmc.ncbi.nlm.nih.gov/articles/PMC4383059/
Source type	Journal Article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	<p>GWAS is used to find interactive factors in the genome that may affect certain traits. The problem? The genome is extremely long and unfeasible for most regular computers to compute. That's why high power computing (HPC) is often used to support GWAS.</p> <p>This article from 2015 discusses the recent advances in data analysis and computing technology to determine that GWAS is more feasible today than ever, and introduces their own framework for supporting HPC in genomic analysis. Despite this, it has been nearly 9 years and technology moves extremely fast. Especially with the exponentially fast development of the graphics cards/AI training industry, it's likely that today's results will be much faster.</p> <p>This study estimates that, in 2015, the "Avoca" IBM Blue Gene/Q supercomputer (from 2011) could render 1.1 million SNP GWAS associations in 5.8 years, multiple times faster than other speed estimates for those which don't utilize this study's framework.</p>
Research Question/Problem/Need	How will high power computing/advances in computing technology and optimization algorithms affect the speed of genomic analysis?

Important Figures

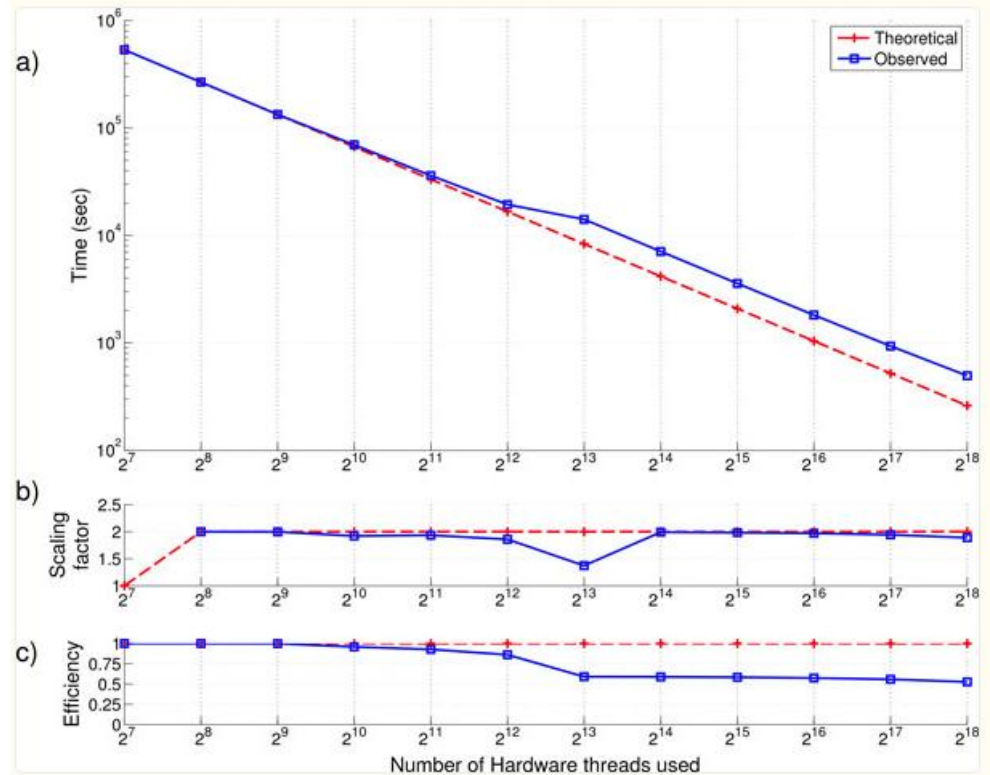


Figure 1. This figure explores the effect on runtime performance as more hardware threads are used with this study's framework. A relatively large number of threads (2^{11}) can be used before noting efficiency dips, and generally, the time decreases at a linear rate as more threads are introduced.

VOCAB: (w/definition)

Runtime performance – How long a program takes to accomplish its task
 HPC – High Power Computing, refers to large servers/supercomputers that handle storing/analyzing large amounts of data

Cited references to follow up on

<https://linkinghub.elsevier.com/retrieve/pii/S0002929710003782>

Follow up Questions

1. What optimization methods have been made to make this computation faster?
2. Which analysis methods were used to estimate the time taken on the computer?
3. How much more effective is it today than from 2015?

Article #18: An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies

Source Title	An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies
Source citation (APA Format)	Liang, X., Wang, Z., Sha, Q., & Zhang, S. (2016). An Adaptive Fisher's combination method for joint analysis of multiple phenotypes in association studies. <i>Scientific Reports</i> , 6(1). https://doi.org/10.1038/srep34323
Original URL	https://www.nature.com/articles/srep34323
Source type	Journal Article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	<p>Many GWAS studies so far have involved looking at the correlation between the genome and one specific phenotype. However, the truth is that among each gene that may have a correlation, they can actually have many different phenotypic correlations that may contribute to a single, complex disease. This study uses a modified Fisher's Method (which is traditionally used to unify p-values for an entire experiment amongst smaller sub-parts with their own respective p-values) as an Adaptive Fisher's Combination Method that does joint analysis involving multiple phenotypes using data from GWAS studies that previously only involved single phenotype correlation.</p> <p>This study also compares this new AFC method to other methods of accomplishing the same task, such as TATES, Tippett's method, Fisher's combination test, MANOVA, MultiPhen, and SUMSCORE. It uses simulations to prove that AFM more often avoids Type I error (the chance of falsely claiming a significant p-value).</p>
Research Question/Problem/Need	How can multiple, single-phenotype GWAS be combined to represent broader results and correlations across many phenotypes?

Important Figures	<div data-bbox="527 205 1494 766" data-label="Figure"> <p>Scenario 2</p> <table border="1"> <caption>Estimated Power Values for Scenario 2</caption> <thead> <tr> <th>Beta</th> <th>AFC</th> <th>TATES</th> <th>Tippett</th> <th>FC</th> <th>MANOVA</th> <th>MultiPhen</th> <th>SUMSCORE</th> </tr> </thead> <tbody> <tr> <td>0.25</td> <td>0.25</td> <td>0.10</td> <td>0.10</td> <td>0.15</td> <td>0.05</td> <td>0.05</td> <td>0.20</td> </tr> <tr> <td>0.30</td> <td>0.65</td> <td>0.30</td> <td>0.30</td> <td>0.45</td> <td>0.20</td> <td>0.20</td> <td>0.55</td> </tr> <tr> <td>0.35</td> <td>0.85</td> <td>0.60</td> <td>0.60</td> <td>0.75</td> <td>0.45</td> <td>0.45</td> <td>0.80</td> </tr> <tr> <td>0.40</td> <td>0.95</td> <td>0.85</td> <td>0.85</td> <td>0.90</td> <td>0.75</td> <td>0.75</td> <td>0.95</td> </tr> <tr> <td>0.45</td> <td>0.98</td> <td>0.95</td> <td>0.95</td> <td>0.98</td> <td>0.90</td> <td>0.90</td> <td>0.98</td> </tr> </tbody> </table> </div> <p>Figure 1. This figure compares the AFC method to other methods that aim to consolidate p-values. The x-axis (represented by Beta) represents the signal strength, or the effect size of the correlation between the SNPs and the phenotypes. In all scenarios, AFC is either dramatically better than most other methods or is on nearly equal footing, finding the most significant advantages at the low signal strength end (which is important because most genomic data contains weaker signals on average).</p>	Beta	AFC	TATES	Tippett	FC	MANOVA	MultiPhen	SUMSCORE	0.25	0.25	0.10	0.10	0.15	0.05	0.05	0.20	0.30	0.65	0.30	0.30	0.45	0.20	0.20	0.55	0.35	0.85	0.60	0.60	0.75	0.45	0.45	0.80	0.40	0.95	0.85	0.85	0.90	0.75	0.75	0.95	0.45	0.98	0.95	0.95	0.98	0.90	0.90	0.98
Beta	AFC	TATES	Tippett	FC	MANOVA	MultiPhen	SUMSCORE																																										
0.25	0.25	0.10	0.10	0.15	0.05	0.05	0.20																																										
0.30	0.65	0.30	0.30	0.45	0.20	0.20	0.55																																										
0.35	0.85	0.60	0.60	0.75	0.45	0.45	0.80																																										
0.40	0.95	0.85	0.85	0.90	0.75	0.75	0.95																																										
0.45	0.98	0.95	0.95	0.98	0.90	0.90	0.98																																										
VOCAB: (w/definition)	None.																																																
Cited references to follow up on	https://pubs.rsna.org/doi/10.1148/radiol.11110173																																																
Follow up Questions	<ol style="list-style-type: none"> 1. How would a method like this perform with actual genomic data? 2. How do methods like this improve the identification of crucial genes with cancer? 3. Are there certain use cases in which methods that work with stronger correlations may be better? In this method still effective? 																																																

Article #19: Detecting Differentially Expressed Genes with RNA-seq Data Using Backward Selection to Account for the Effects of Relevant Covariates

Source Title	Detecting Differentially Expressed Genes with RNA-seq Data Using Backward Selection to Account for the Effects of Relevant Covariates
Source citation (APA Format)	Nguyen, Y., Nettleton, D., Liu, H., & Tuggle, C. K. (2015). Detecting Differentially Expressed Genes with RNA-seq Data Using Backward Selection to Account for the Effects of Relevant Covariates. <i>Journal of agricultural, biological, and environmental statistics</i> , 20(4), 577–597. https://doi.org/10.1007/s13253-015-0226-1
Original URL	https://pmc.ncbi.nlm.nih.gov/articles/PMC4666287/
Source type	Journal Article
Keywords	False discovery rate, Generalized linear model, Quasi-likelihood, Residual feed intake
#Tags	
Summary of key points + notes (include methodology)	<p>Linear regression is commonly used in GWAS studies to analyze data and identify correlations between SNPs and phenotypes. However, for researchers who want to know even more about the direct effects of the proteins that cells make and their effects on phenotypes and more, it would be more worth to do an RNA-Sequencing study that focuses on gene expression.</p> <p>In using regression methods to analyze RNA-Seq data, the challenge comes in identifying differentially expressed genes, genes that are expressed differently based on some other environmental factor. This study proposes a backward selection strategy to select a set of covariates that can be assigned to effects to account for their variation when searching for differentially expressed genes.</p> <p>They then use a simulation to show that the backward selection procedure works better than other methods.</p>
Research Question/Problem/Need	How do we use linear regression in RNA-Sequencing analysis to more reliably determine differentially expressed genes and how their expression changes based on environmental factors?
Important Figures	None.
VOCAB: (w/definition)	Differentially expressed genes – Genes that change their expression based on environmental factors

Cited references to follow up on	https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25
Follow up Questions	<ol style="list-style-type: none">1. Is this the only use of regression in RNA-Sequencing analysis?2. Could this be generally applied to the broader scope of GWAS to consider signals that are affected by environmental factors?3. Are alternative regressions more optimal than linear regression?

Article #20: Regularization and Variable Selection Via the Elastic Net

Source Title	Regularization and Variable Selection Via the Elastic Net
Source citation (APA Format)	Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. <i>Journal of the Royal Statistical Society. Series B, Statistical Methodology</i> , 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
Original URL	https://academic.oup.com/jrsssb/article/67/2/301/7109482
Source type	Journal Article
Keywords	
#Tags	
Summary of key points + notes (include methodology)	The Elastic Net (commonly abbreviated as Enet) is commonly used in similar regression scenarios to the LASSO (particularly variable selection as it pertains to genomic analysis). This study introduces this method and finds that it is much better particularly for sparse datasets. Additionally, it also introduces a LARS-EN method to compute the regularization efficiently, similar to the LARS algorithm for the LASSO, which also optimizes runtime performance.
Research Question/Problem/Need	How do we perform more accurate variable selection than the LASSO?

Important Figures

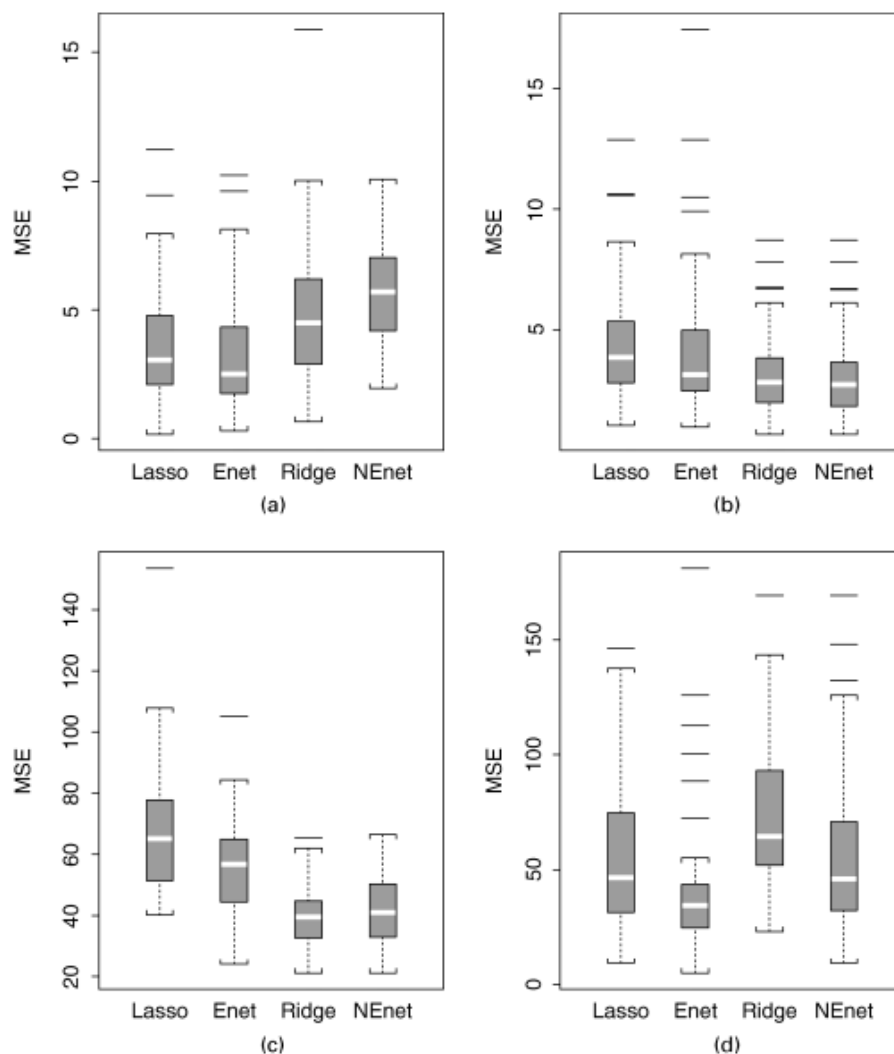


Figure 1. These box and whisker plots display the performance of four different regression algorithms across four different experiments. Particularly, it is comparing their mean squared error (MSE). Note that Enet, in every single example, outperforms the LASSO and especially improves on its performance in experiments c and d.

VOCAB: (w/definition)

None.

Cited references to follow up on

<https://projecteuclid.org/journals/annals-of-statistics/volume-32/issue-2/Least-angle-regression/10.1214/009053604000000067.full>

Follow up Questions

1. How does its runtime performance compare to LASSO?
2. Do more current methods exist that accomplish even better regression?
3. Why is LASSO used more than Enet in modern genetic studies?