Bioinformatics Program
Technical University of Munich
Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Search for evidence of recombination in Alternaria solani using genomic tools

Alexander Fastner

Bioinformatics Program
Technical University of Munich
Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Search for evidence of recombination in Alternaria solani using genomic tools

Suche nach Hinweisen auf eine Rekombination in Alternaria solani mit genomischen Werkzeugen

Author: Alexander Fastner
Supervisor: Dr. Remco Stam
Advisors: Prof. Dr. Ralph Hückelhoven, Chair of Phytopathology
Prof. Dr. Dimitri Frischmann, professorship Bioinformatics
Submitted: 13/10/2021

Declaration of Authorship

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Date 13/10/2021 Alexander Fastner

Zusammenfassung

Der Pilzerreger *Alternaria solani* ist dafür bekannt, bei Tomaten und Kartoffeln die Krautfäule zu verursachen. Ein Befall mit *A. solani* kann zu massiven Ertragseinbußen führen und ist die größte wirtschaftliche Schadensquelle für Tomatenernten in tropischen und subtropischen Regionen. Dies wird durch eine Vielzahl von Fungiziden bekämpft, von denen ein gängiges auf Succinat-Dehydrogenase-Inhibitoren (SDHI) abzielt. Die meisten *A. solani* Stämme weisen eine gewisse Resistenz gegen diese Fungizide auf. Jedoch ist die Resistenz gegen diese Fungizide in den letzten Jahren in Europa schneller als erwartet gestiegen. Darüber hinaus sind die verantwortlichen Mutationen in mehreren Populationen in ganz Europa aufgetreten. *A. solani* gehört zur Gruppe der sogenannten Fungi imperfecti, die keine bekannte sexuelle Fortpflanzungsphase haben. Eine mögliche Ursache für die schneller als erwartete Verbreitung und Annahme dieser Fungizidresistenz verleihenden Mutationen kann das Auftreten einer sexuellen Rekombination sein. Sollte dies der Fall sein waeren neue Ansaetze bei der Bekaempfung moeglich
Ausgehend von FASTA-Dateien aus 48 Proben aus ganz Europa teilen wir diese nach Chromosomen auf und erzeugen Dendrogramme mit RAxML als Eingabe für unsere Tools.
Wir verwenden die Rekombinationserkennungstools clonalFrameML und LDHelmet, um beide mögliche Rekombinationsereignisse im Genom von *A. solani* zu identifizieren. Wir führen clonaleFrameML mit alternativen Eingabeparametern wie mit dem (transition/transversion) aus, um den Einfluss der Anfangsparameter auf die Ausgabe zu bestimmen und die Robustheit unserer Ergebnisse zu überprüfen. Anschließend filtern und visualisieren wir unsere Ergebnisse in R mit ggplot2. Der Vergleich der Ergebnisse dieser Tools zeigt eine Überlappung der Vorhersagen, die darauf hindeuten, dass irgendwann ein Rekombinationsereignis eingetreten ist. Die Bestimmung, ob es sich bei den erkannten Regionen um Rekombinationsereignisse der Vorfahren handelt, ist nicht Gegenstand dieser Arbeit und kann Gegenstand einer zukünftigen Untersuchung sein.

Abstract

The fungal pathogen *Alternaria solani* is known to cause early blight in tomatoes and potatoes. An *A. solani* infestation can lead to massive losses in crop yield and is the largest source of economic damage to tomato harvests in tropical and subtropical regions. This is combated by a variety of fungicides, a common one targeting succinate dehydrogenase inhibitors (SDHI). Most strains of *A. solani* have some resistance to these fungicides. However in recent years in Europe the resistance to these has increased at a faster than expected rate. In addition the responsible mutations have arisen in several populations across Europe. *A. solani* belongs to the group known as *fungi imperfecti* which do not have a known sexual reproductive phase. A possible cause for the faster than anticipated spread and adoption of these fungicide resistance granting mutations may be the occurrence of sexual recombination. Should this be the case new options to combat this would be made available.
Starting with FASTA files from 48 samples from across Europe we split these by chromosome and generate dendrograms with RAxML as input to our tools.
We utilize the recombination detection tools clonalFrameML and LDHelmet to both identify possible recombination events in the *A. solani* genome. We run clonalFrameML with alternate input parameters such as transition/transversion ratio, to determine the impact of initial parameters on the output and verify the robustness of our results.
Then we filter and visualize our results in R using ggplot2. The comparison of results from these tools shows an overlap of predictions which suggests a recombination event occurring at some point. Determining whether the detected regions are ancestral recombination events is out of scope for this paper and can be subject of a future investigation.

Acknowledgements

I'd like to thank my Supervisor Remco Stam for providing this real and interesting use case as well as his guidance and support.

A special thanks to Drazen Jalsovec for helping me get set up on the server.

Table of Contents

Introduction 1

Biological background 1.1
Alternaria solani is the fungal pathogen that causes early blight in tomatoes and potatoes. On tomatoes it starts as a small brown/black ring on the leaves that can slowly spread to the stem and to the fruit. It is a necrotrophic pathogen that kills host tissue by breaking down the cell wall and feeding on the inside cell material. *Alternaria solani* has been known as a member of the *fungi imperfecti*; it has not been observed to have a sexual reproductive phase. “Approximately one fifth of all described fungal species have no known sexual stages. In ascomycete fungi, asexuality has evolved independently many times from sexual ancestors” (**Milgroom MG, Jiménez-Gasco M del M 2014**). Instead it reproduces asexually through mitosis by means of conidia. A fully mature lesion is covered in dark dusty spores which can overwinter either on the seeds or crop residue. These can survive for extended periods of time in the ground, and it is therefore recommended to alternate to a non-Solanaceous crop for three years after a harvest. Solanaceous crops or Nightshades are a large family of plants that includes crops such as potatoes, tomatoes and bell peppers.

Motivation 1.2
An uncontrolled *Alternaria solani* outbreak can result in yield losses of 40%. (**Leiminger JH, Hausladen H**. **2012**) In some instances annual economic yield loss has even been estimated at 79%. (**Adhikari P, Oh Y, Panthee DR**. **2017**) *A. solani* spreads most quickly under warm, moist conditions to infect plants. Methods such as using drip irrigation and watering early in the morning have shown to be effective at reducing the infection. (**McGovern,Dr Robert, Hall F**. DISEASE MANAGEMENT: Early Blight of Tomato) In this study we attempt to find evidence for sexual recombination in *A. solani*. This is important because it might prove helpful when designing more cost-effective integrated pest management systems. Dealing with a low spread low mutation rate pathogen allows for use of other strategies as when dealing with a sexual pathogen with more rapid and broad spread.

Fungicides 1.3
Fungicides are used to reduce the spread and impact of *Alternaria solani* on crop yields. The main types of fungicide used are succinate dehydrogenase inhibitors (SDHI). Succinate dehydrogenase is an integral part of the mitochondrial respiration chain known as complex II. (**FRAC HOME**: FUNGICIDE RESISTANCE ACTION COMMITTEE) Many strains of *Alternaria solani* show some level of resistance to popular SDHI’s but recently strains with high SDHI resistance have been found in Germany. The genetic mutations responsible for this seem to have arisen separately throughout Europe as they are not geographically clustered. (**Einspanier S, Susanto T, Metz N, Wolters PJ, Vleeshouwers VGAA, Lankinen Å, Liljeroth E, Landschoot S, Ivanović Ž, Hückelhoven R, *et al.* 2021**). Previous work shows that samples obtained in a similar geographical location do not show as high a similarity as one might expect, and these genetic differences could be explained by occurrences of sexual recombination.

Dataset 1.4
48 *Alternaria solani* samples were collected from several locations in several countries across europe. These samples are from the following regions: Bavaria, Lower Saxony, Serbia, Belgium, southern Sweden and 4 samples from the United States.

Tools 1.5
Here we utilize two recombination detection tools ClonalFrameML (**Didelot X, Wilson DJ**. **2015**) and LDHelmet (**Chan AH, Jenkins PA, Song YS**. **2012**) to support the theory that there is or was recombination in strains of *Alternaria solani.*
ClonalFrameML is a software package used to perform inference of recombination in bacterial genomes. The program uses both sequence data as well as an ancestry tree as input.
LDHelmet is a tool used to find fine-scale evidence for recombination. Its primary advantage is its resistance to noise in the data, which comes at the cost of a longer runtime and lack of an integrated visualization tool. In this paper both tools are utilized and the data compared to judge the robustness of the given results.

Method 2

Tool research 2.1

There are many existing tools to determine the likelihood of genetic differences being caused by recombination. Many of these have parameters aside from just sequence information that play a role in that algorithm. Programs designed to work with a specific number of chromosomes, haploid/diploid organisms, bacteria, or known examples of recombination on reference genomes . Additionally finding a tool to detect recombination in an organism with no known sexual reproductive phase proved challenging. As a starting point clonalFrameML is an existing tool which was designed precisely to detect occurrences of recombination. While initially developed for use on Bacteria aside from longer runtimes it should achieve the same function. LDHelmet is another tool developed to detect recombination with the primary advantage of being noise resistant. Comparing Data from both tools allows us to judge the robustness of our results and to take a closer look at any overlaps.

Fungal Material 2.2

The dataset used was provided by Dr. Remco Stam (LS Phytopathology, TUM). In 5 distinct localities across Europe 48 isolates were sampled from different fields. Symptomatic leaves were taken and dried between sheets of tissue paper and sterilized before being placed on SNA (0.2 g/l glucose, 0.2 g/l sucrose, 0.5 g/l MgSO4-7H2O, 0.5 g/l KCl, 1.0 g/l KH2PO4, 1.0 g/l KNO3, 22.0 g/l agar; 600 µl/l 1M NaOH). From each of the samples one spore was taken to be further propagated. (**Einspanier S, Susanto T, Metz N, Wolters PJ, Vleeshouwers VGAA, Lankinen Å, Liljeroth E, Landschoot S, Ivanović Ž, Hückelhoven R, *et al.* 2021**)

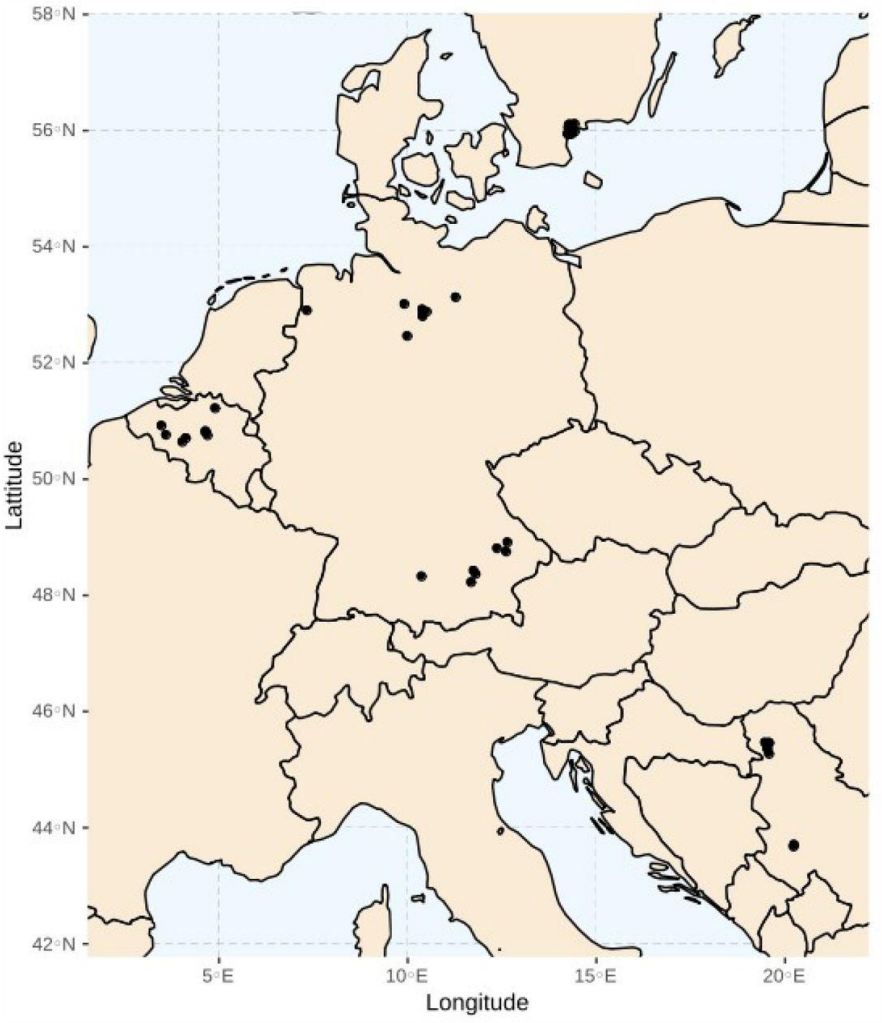


Figure 1
(Map of the sampling locations of the *A. solani* isolates) Einspanier S, Susanto T, Metz N, Wolters PJ, Vleeshouwers VGAA, Lankinen Å, Liljeroth E, Landschoot S, Ivanović Ž, Hückelhoven R, *et al.* 2021

DNA extraction 2.3

The samples studied in this paper were previously gathered, sequenced and mapped. The DNA was extracted using a two-day chloroform:isoamyl alcohol-based protocol as described in (**Stam R, Einspanier S, Susanto T. 2021**).

Mapping 2.4

The mapping described below was done by another team. (**Einspanier S, Susanto T, Metz N, Wolters PJ, Vleeshouwers VGAA, Lankinen Å, Liljeroth E, Landschoot S, Ivanović Ž, Hückelhoven R, *et al.* 2021**) However it is important to understand the mapping and how this data was gathered.

The mapping of all samples to the reference genome NL03003 (**Wolters et al., 2018**) was done with the Burrows-Wheeler alignment tool (**Li & Durbin, 2009**). The quality of the raw sequence data was checked using FastQC. The reads deduplication was done with the picard tool Markduplicate. The Single-nucleotide polymorphisms (SNPs) were called using GATKs Haplotype-Caller in GVCF mode (**McKenna et al., 2010**) with default settings, except for having ploidy set to one. SNPs were filtered out using GATK with the following parameters: low mapping quality rank sum test (MQRankSum < - 12.5), low quality by depth (QD < 2.0), low read pos rank sum test (ReadPosRankSum < −8.0), high Fisher strand difference (FS > 60.0), low RMS mapping quality (MQ < 40.0), high strand odds ratio (SOR > 3.0), high haplotype score (HaplotypeScore > 13.0). Insertion or deletions (indels) were filtered out using GATK with these parameters: low quality depth (QD < 2.0), low read pos rank sum test (ReadPosRankSum < -20.0), and high Fisher strand difference (FS > 200.0). Afterwards, SNP clusters and SNPs close to indels were removed using the SnpSift filter (**Cingolani et al., 2012a**). Subsequently the summary of variant analysis was generated by SnpEff (**Cingolani et al., 2012b**).

Input Data processing 2.5

The initial data consisting of complete genome FASTA files from 48 samples was downloaded from a repository by Stam Lab. In order to run LDHelmet/clonalFrameML additional files containing ancestral relationships were needed. To run clonalFrameML one needs a FASTA file and a tree in Newick format. The Newick format is a way to represent a tree with edge length using parentheses and commas. These Newick trees were generated using the program Randomized Accelerated Maximum Likelihood (RAxML) (**A. Stamatakis 2014**.). For the initial trial run the full genome FASTA was used. Afterwards to achieve a greater resolution each sample genome was split into all 10 individual chromosomes then reassembled into 10 files with the respective chromosome from each sample. This was achieved with several custom bash scripts. The headers and file names were then adjusted as clonalFrameML requires branch names in the tree to match headers in the Fasta. These new FASTAs were then used to generate new RAxML files to be used as input to clonalFrameML.
LDHelmet, being more compute intensive than clonalFrameML, needed to be run on a server in Stam Lab. Running LDHelmet only required the FASTA files created earlier.

ClonalFrameML 2.6

ClonalFrameML was published in 2015 by Xavier Didelot and Daniel Wilson as a software package to perform inference of recombination in bacterial genomes. (**Didelot X, Wilson DJ**. **2015**) As inputs the program takes a starting tree in Newick format and an alignment in FASTA or XMFA. The number of leaves in the tree must match the number of sequences in the alignment. A homoplasy is a mutation which occurs on multiple branches. ClonalFrameML creates an ML tree and calculates the expected mutations and transitions/transversions. Then it counts homoplasies and compares that distribution to what it calculated to classify what to mark as likely recombination.

To start, ClonalFrameML creates a position cross reference file to be able to lookup the position of a node in the given tree in the sequence file. Then the ancestral sequences are reconstructed by use of maximum likelihood (ML). Next a Baum-Welch Expectation-Maximization algorithm is used to get estimates of the recombination parameters, and the branch lengths. Baum-Welch is a special case of typical Expectation Maximization (EM) as it utilizes the forward-backward algorithm to find hidden parameters in the Hidden Markov Model (HMM) in the E step of EM. The next step is to construct a file of recombination events called importation status that is inferred by the Viterbi algorithm. The Viterbi algorithm finds the most likely sequence of observed events in a HMM. The uncertainties/error bars are calculated for the above steps by drawing random samples with replacement from the dataset to calculate sample distribution statistics in the process called bootstrapping.

ClonalFrameML also includes an R script with which to create a graphical output if the correct R packages *Ape* and *Phangorn* are installed as well.
The programmers of clonalFrameML did not parallelize their algorithm as it was likely not necessary for the short bacterial Genomes they designed it for. Running their tool on fungal genomes was possible on a desktop but did take over 12 hours to run on average.

LDhelmet 2.7

LDHelmet is a tool developed to find fine-scale recombination rates from genetic data. (**Chan AH, Jenkins PA, Song YS**. **2012**) The main inputs to this software are a FASTA or FASTQ file, as well as a mutation transition matrix, if not provided the default is the Jukes-Cantor mutation matrix. The Jukes-Cantor mutation matrix assumes equal base frequencies and mutation rates. This program makes several major assumptions: That the DNA sequences are randomly drawn from a single population, that the population follows neutral evolution and has a constant size, and a constant recombination rate across the sequence.
Simply put LDHelmet determines what to classify as recombination by creating an estimated recombination map as a prior which is stored as a step-wise function for each haplotype. A ML is calculated for each of these maps and compared to the estimated recombination map and determines differences of 1 order of magnitude or more.

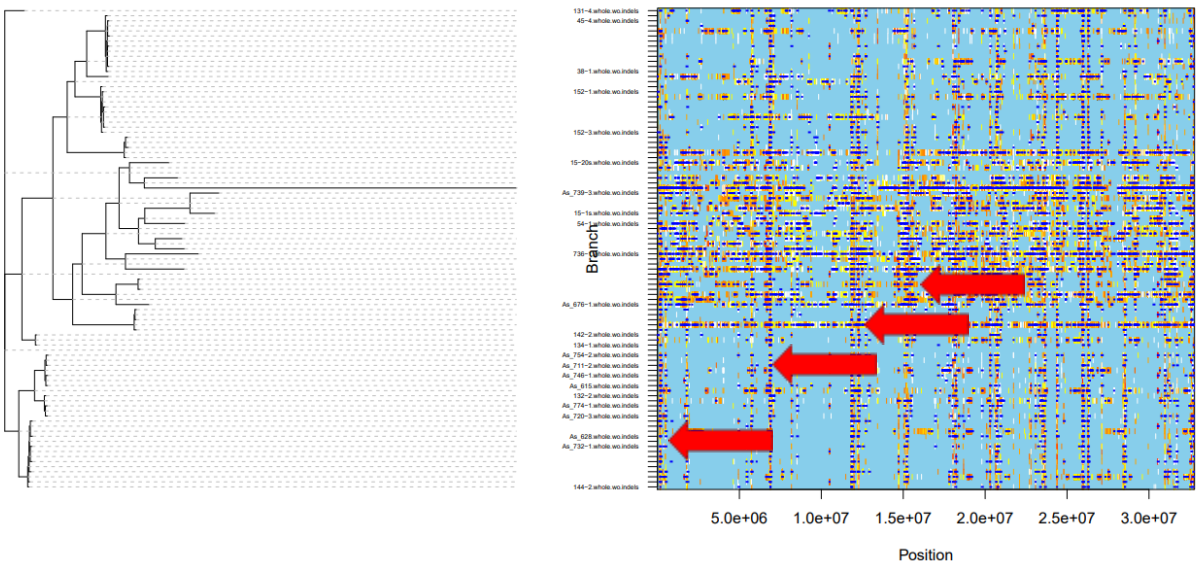
The program first creates a haplotype configuration file (with the .conf extension) which records the various haplotypes from the FASTA sequence. One can set the window size for calculating the haplotype configuration file to include a certain number of SNPs which in this case was the recommended 50. Then it creates a likelihood lookup table (.lk) from that configuration file and several parameters. Θ is the population-scaled mutation rate, in units of 1/bp. The ρ is a series of values and increments in between those values used to generate a grid of population scaled recombination rates. The authors of LDhelmet also recommend using Pad´e coefficients. Pad´e coefficients are generated using an asymptotic sampling formula from the haplotype configuration file and are in essence an approximation of the given function. Any number of these can be generated from the Θ and more will improve accuracy at the cost of computation time.

Now that all the prerequisite files have been created the main algorithm, reversible-jump Markov Chain Monte Carlo (rjMCMC) is performed. Ancestral allele priors indicate the probability of each allele being ancestral for every SNP. The results are then post processed to show the mean, percentiles, and other specifiable information. The program is run once for the combined Fasta including all 10 chromosomes then again for each chromosome individually.
The output from LDHelmet still needed to be visualized which was done using a custom R script using ggplot2 (**Wickham, H. 2016**). The resulting plots show the mean rho/bp values plotted along the positions of the sequence along the chromosomes or complete genome.

Results 3

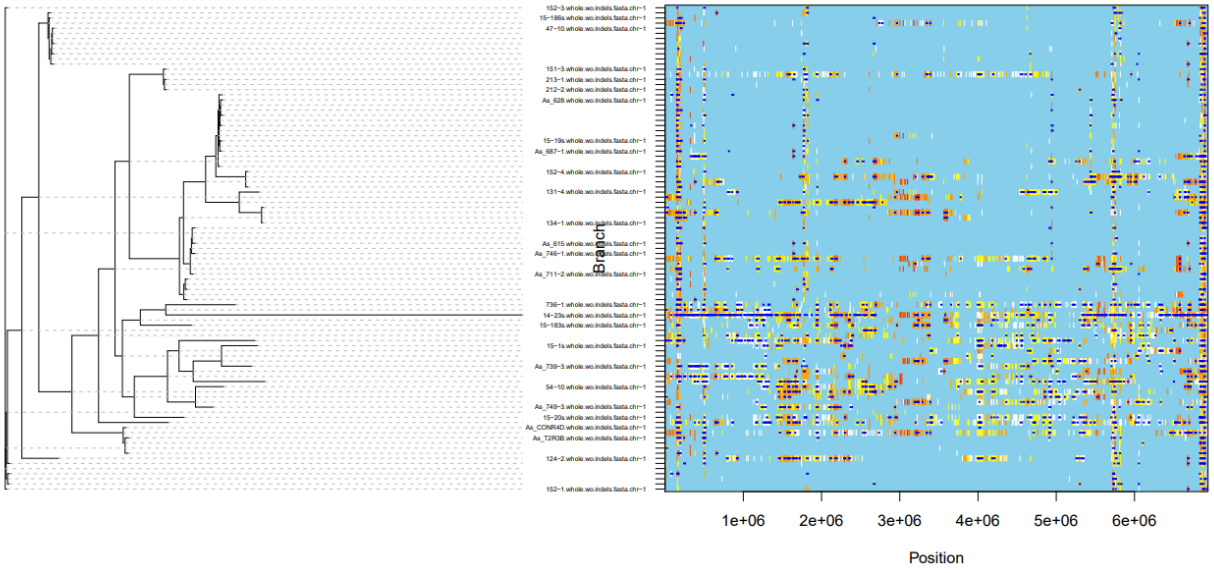
ClonalFrameML 3.1

Figure 2 shows the output of ClonalFrameML run on the combined genome of *Alternaria solani.*
On the left-hand side is a dendogram with the corresponding sample names in the middle. The rows on the right-hand side reflect clonalFrameML's prediction for likelihood of recombination for each position across the genome. The output shown here is color coded. Dark blue - detected recombination :light blue, yellow, orange, red - Homoplasies or raw mutations. As mentioned in methods a Homoplasy is a mutation that occurs on multiple branches. Here these colors are a scale from light blue to red showing how many times a mutation occurs on various branches. These homolasies are what the program finds and then uses to predict the likelihood of these mutations being due to recombination.
The continuous blue horizontal lines are likely an artifact of the way ClonalFrameML works as they only appear on lines that aren't labeled. The non-labeled rows are genomes generated by the program as a stand in common ancestor link to other samples. This could indicate that the sample size may be too diverse for analysis of two branches that are labeled near each other. ClonalFrameML may think the lineages are so diverse as to require recombination when in reality there is merely not enough data connecting those samples. Predictions of recombination in these non-existent strands could also just be an artifact of the program.
Strong evidence for recombination would be the same positions marked in blue across several samples. This would be indicated by blue indicators spread vertically in these figures.



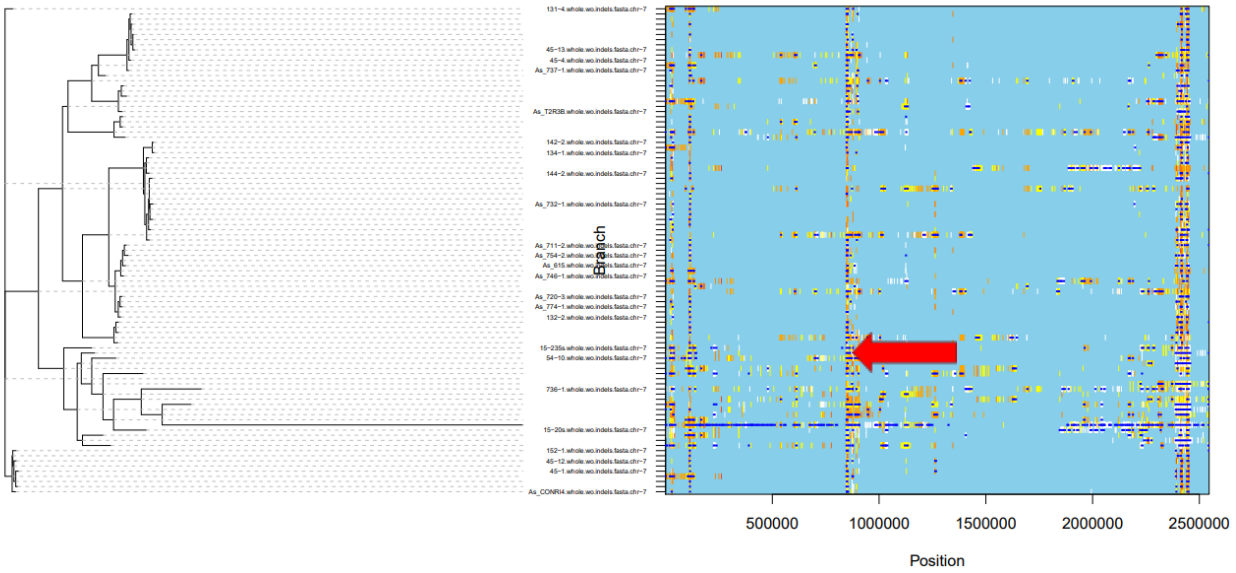
(Figure 2)
Whole Genome clonalFrameML
On the left-hand side is a dendogram with the corresponding sample names in the middle. The rows on the right-hand side reflect clonalFrameML's prediction for likelihood of recombination for each position across the genome. Dark blue represents detected recombination. Light blue, yellow, orange and red are the number of detected homoplasies.
Here one can see the output for the combined genomes. Of note are the numerous vertical blue lines which suggest evidence of recombination shared by different samples.

In **Figure 2** one can see several somewhat contiguous vertical blue lines which could indicate recombination, however most of these fall on the edges of chromosomes.
After the initial run it was clear that there were likely some artifacts occurring at the edges of the chromosomes. ClonalFrameML was initially developed to be used on bacterial genomes. As we are using a non-circular genome the program might be detecting repeating sequences on both sides of the chromosome and interpreting these as recombination. On bacteria this would be a smaller jump/cross than on the actual chromosomes used in this paper.
To get a better understanding and resolution the program was then run on all 10 chromosomes individually.

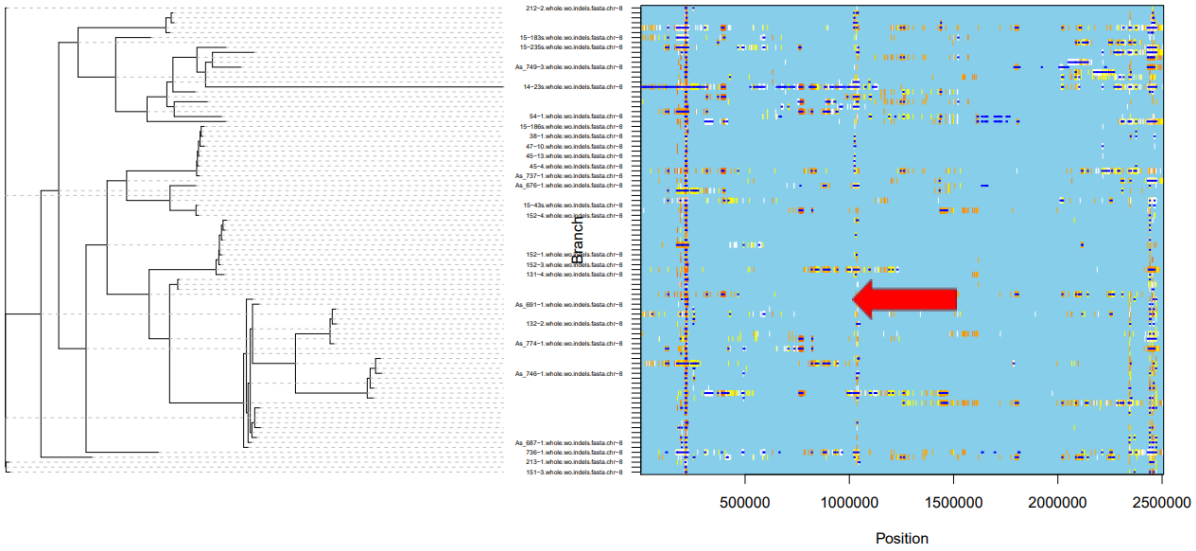


(Figure 3)
Chromosome 1 clonalFrameML
On the left-hand side is a dendogram with the corresponding sample names in the middle. The rows on the right-hand side reflect clonalFrameML's prediction for likelihood of recombination for each position across the genome.
There are many detected homoplasies spread throughout the sequence. Several lines apart from those at the very edges of the chromosome are also visible.

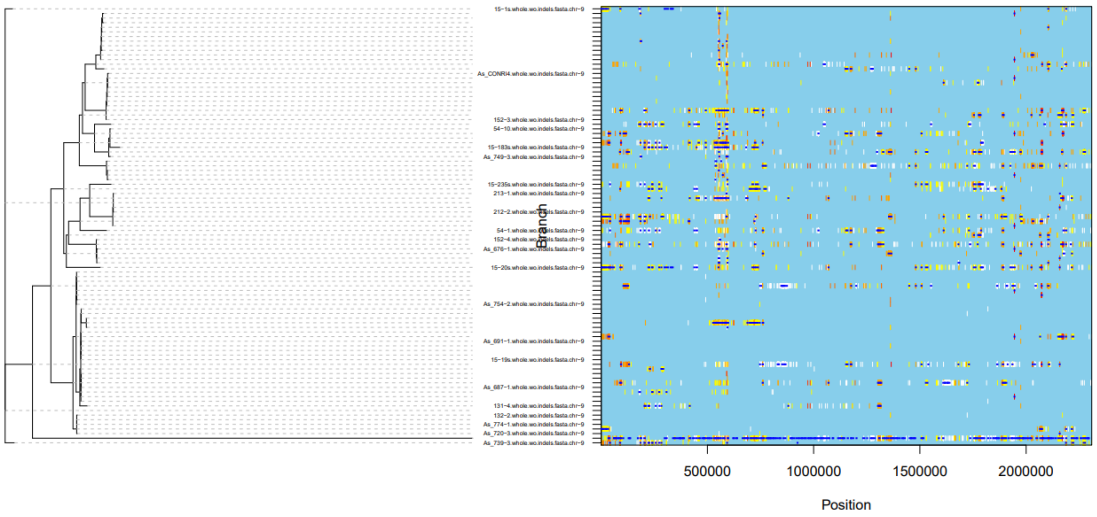
Figure 3 shows the results from chromosome 1 where one can see the lines of dark blue predictions for recombination with the most prominent being at the ends of the chromosome. There was however also evidence for other recombination events throughout the length of the sequence.
Chromosomes where this is especially clear are chromosomes 7, 8 and 9 which are shown in **Figure 4, Figure 5, and Figure 6** respectively. The extremely strong signal in **Figure 4** at ~80,000 could also be attributed to that region being in the centromeric region. This region contains many repeat sections which might confuse the algorithm. However this is something that does not occur in the other chromosomes which supports that this is not the case here. Also of interest is that the predicted recombination takes place in the actual isolates and not the ancestors. This suggests it is unlikely that all these regions would be like this with no recombination event having occurred. The strong signals such as the one at position ~800,000 on chromosome 7, and ~1,100,000 on chromosome 8, 600,000 on chromosome 9 suggest there is more than just noise here.



(Figure 4)
Chromosome 7 clonalFrameML
On the left-hand side is a dendogram with the corresponding sample names in the middle. The rows on the right-hand side reflect clonalFrameML's prediction for likelihood of recombination for each position across the genome.
Of particular note here is the detected recombination at ~800,000.

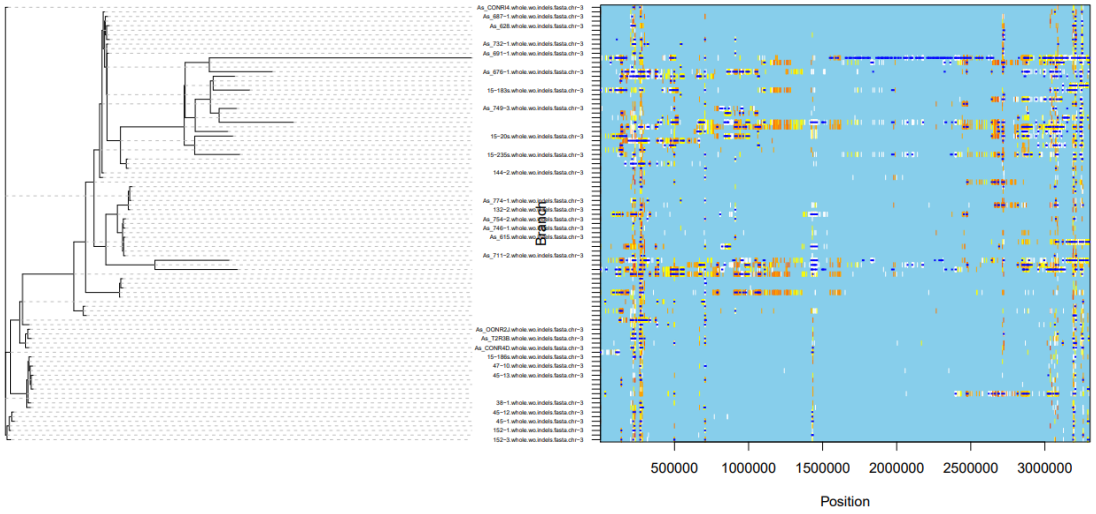


(Figure 5)
Chromosome 8 clonalFrameML
On the left-hand side is a dendogram with the corresponding sample names in the middle. The rows on the right-hand side reflect clonalFrameML's prediction for likelihood of recombination for each position across the genome.
The section at ~1,100,000 suggests a recombination event, whereas the lines at the beginning and towards the end are likely artifacts of the program.



(Figure 6)
Chromosome 9 clonalFrameML
On the left-hand side is a dendogram with the corresponding sample names in the middle. The rows on the right-hand side reflect clonalFrameML's prediction for likelihood of recombination for each position across the genome.
Here there is a strong signal at ~600,000 though there are many detected homoplasies throughout.

While these chromosomes showed stronger signs of recombination, others didn't have as clear a signal. In chromosome 3 **Figure 7** there is less of a clear signal for recombination despite there being many regions where homoplasies were detected across several branches. While not predicted as recombination by clonalFrameML, this shows that many mutations did appear across several branches leading us to take a closer look at this region when examining the results from LDHelmet.



(Figure 7)
Chromosome 3 clonalFrameML
On the left-hand side is a dendogram with the corresponding sample names in the middle. The rows on the right-hand side reflect clonalFrameML's prediction for likelihood of recombination for each position across the genome.
There is less of a clear signal here than some others. Though slight lines are visible at ~700,000 and 1,400,000.

To test the robustness of our results we ran ClonalFrameML several times with altered parameters to measure the impact of starting conditions on the final prediction. The tested parameters are shown in **Table 1** and were run individually and independently. The results from all runs are visually identical and result in no change to the final prediction. This suggests that the predictions by clonalFrameML are not greatly impacted by the parameters chosen and thus that the results are robust.

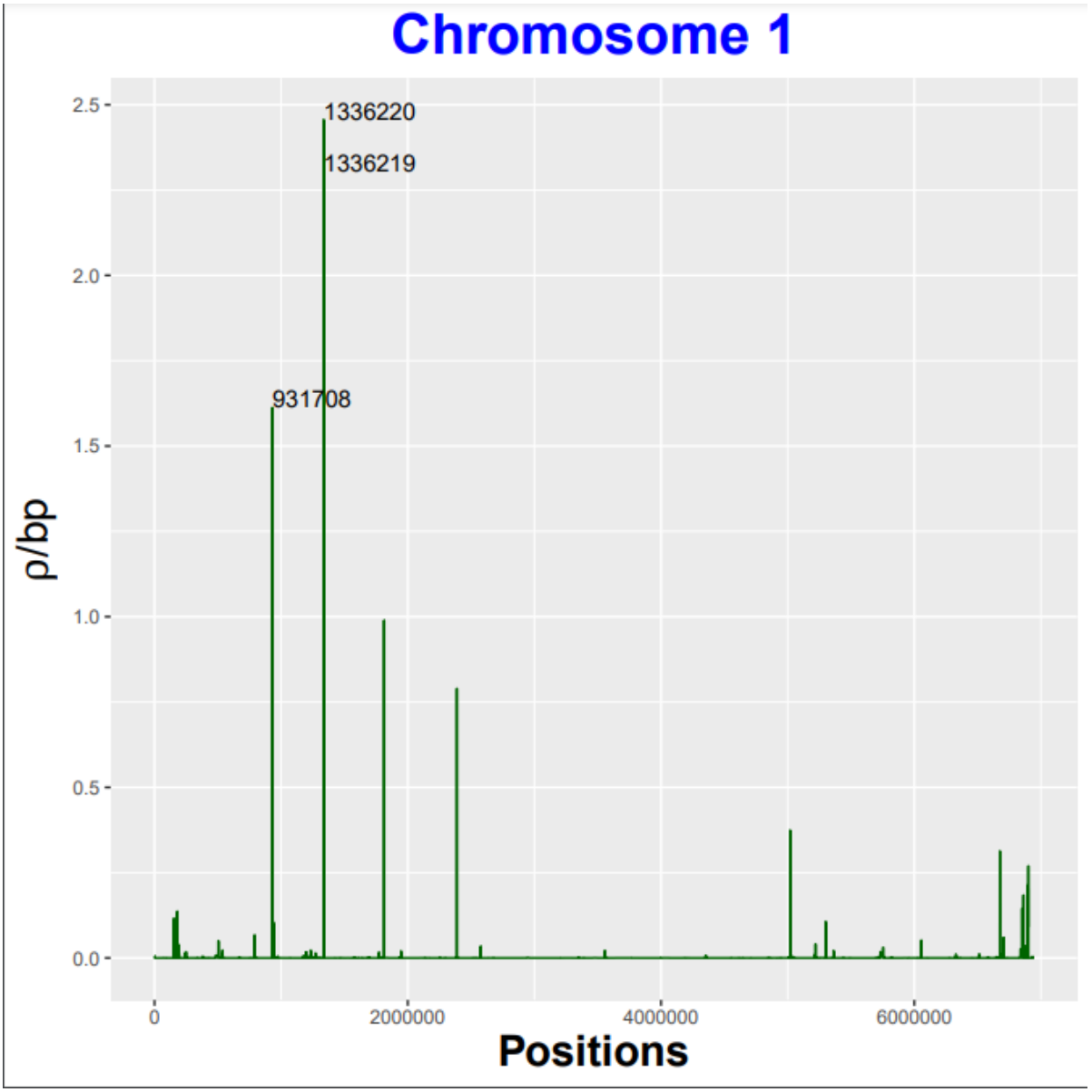
Parameter	Default	Alternate parameters
R/theta (Relative rate of recombination to mutation)	0.1	.01, .001
Kappa (Relative rate of transitions to transversions)	2.0	1.0, 3.0
Nu (Mean divergence of imported DNA)	0.1	1.0, 0.01

(Table 1)
altered parameters
This shows the altered parameters tested to determine their effects on the outcome of clonalFrameML. They were each tested individually and independently of each other.

LDHelmet 3.2

To get further evidence for the recombination detected by clonalFrameML we also used/ran LDHelmet and compared the results. These figures were custom generated utilizing ggplot2 in R. The X axis represents the position in the sequence and the Y shows the recombination rate/base pair (ρ/bp).

In **Figure 8** one can see the visualized results produced by LDHelmet which are less prone to noise and more easily parsable. For chromosome 1 the (ρ/bp) values are rather high with 3 SNPs having a mean value over 1. While simplistic the visualization in R is very clear and easier to interpret than the result offered by clonalFrameML.



(Figure 8)
Chromosome 1 LDHelmet
Shown here is the output of LDHelmet visualized in R. The X axis plots the positions of SNPs. The Y axis represents the recombination rate/base pair (ρ/bp) values. The labeled peaks are the left position of each given SNP with a mean (ρ/bp) greater than 1.
The SNP at 1336219 is one bp and is adjacent to the next which goes from 1336220-1336270.

The clearest results from LDHelmet specifically are seen in chromosomes 8 and 9 in **Figures 9 and 10** respectively. Chromosome 8 shows the biggest mismatch between LDHelmet's prediction and that of clonalFrameML. While both show strong indicators of recombination events they do not line up.
The predicted locations for recombination for chromosome 9 however match up well with the results from clonalFrameML and the high (ρ/bp) values suggest there is quite likely something there. Interestingly the detected recombination is very localized to only one specific region. Also as seen in **Figure 10** neither the clonalFrameML results nor the LDHelmet ones have the detected recombination at either ends of the chromosome. Why this is the case only here is unclear and may contribute to the high (ρ/bp) values seen here.

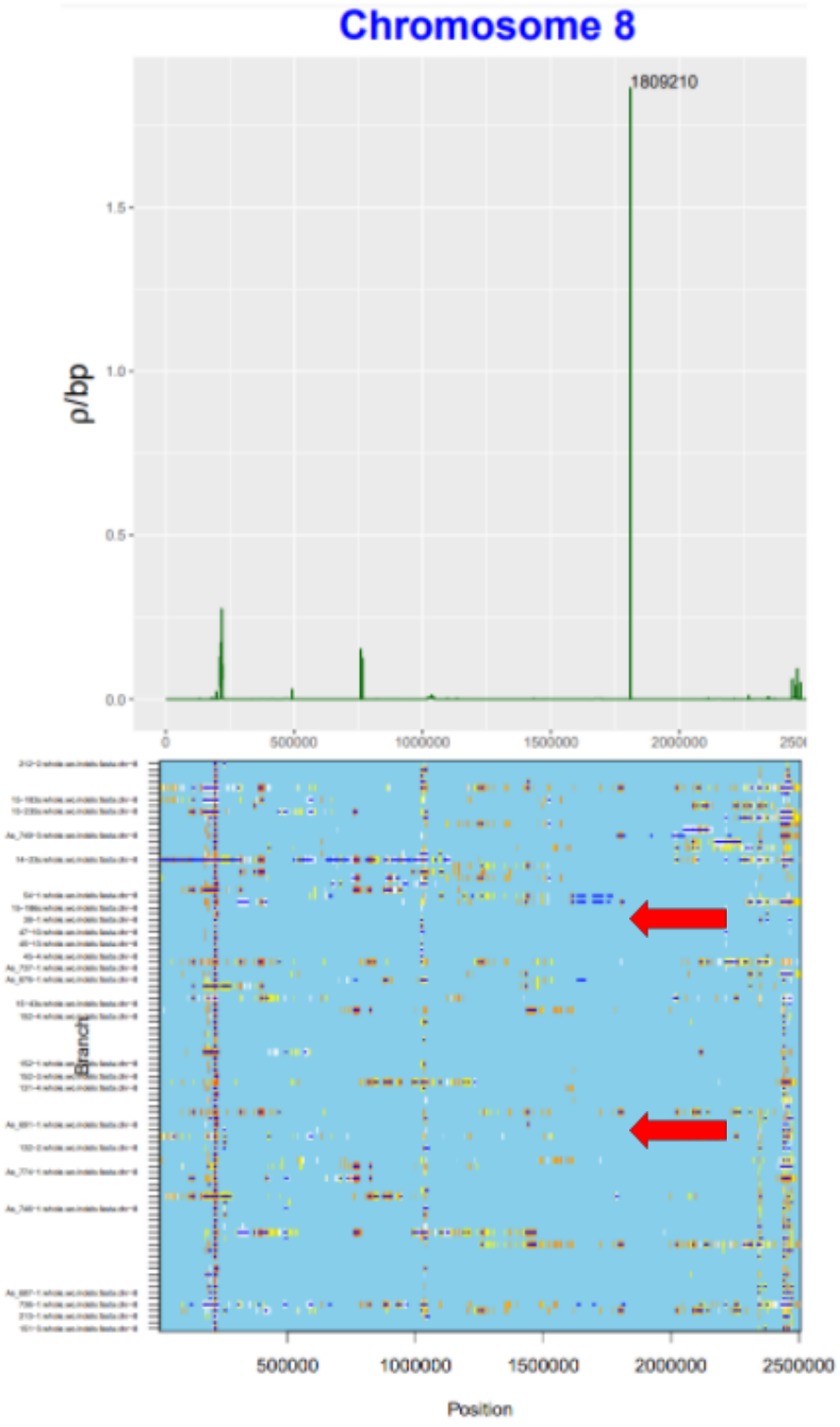


Figure 9
Chromosome 8 comparative clonalframeML LDHelmet
Top panel depicts output of LDHelmet. Lower panel shows results from clonalframeML as described in legend of figure 2.
The prediction by LDHelmet matches up well with that of clonalFrameML with the exception of the major spike at 1,800,000. One would have expected a more visible pattern in clonalframeML.

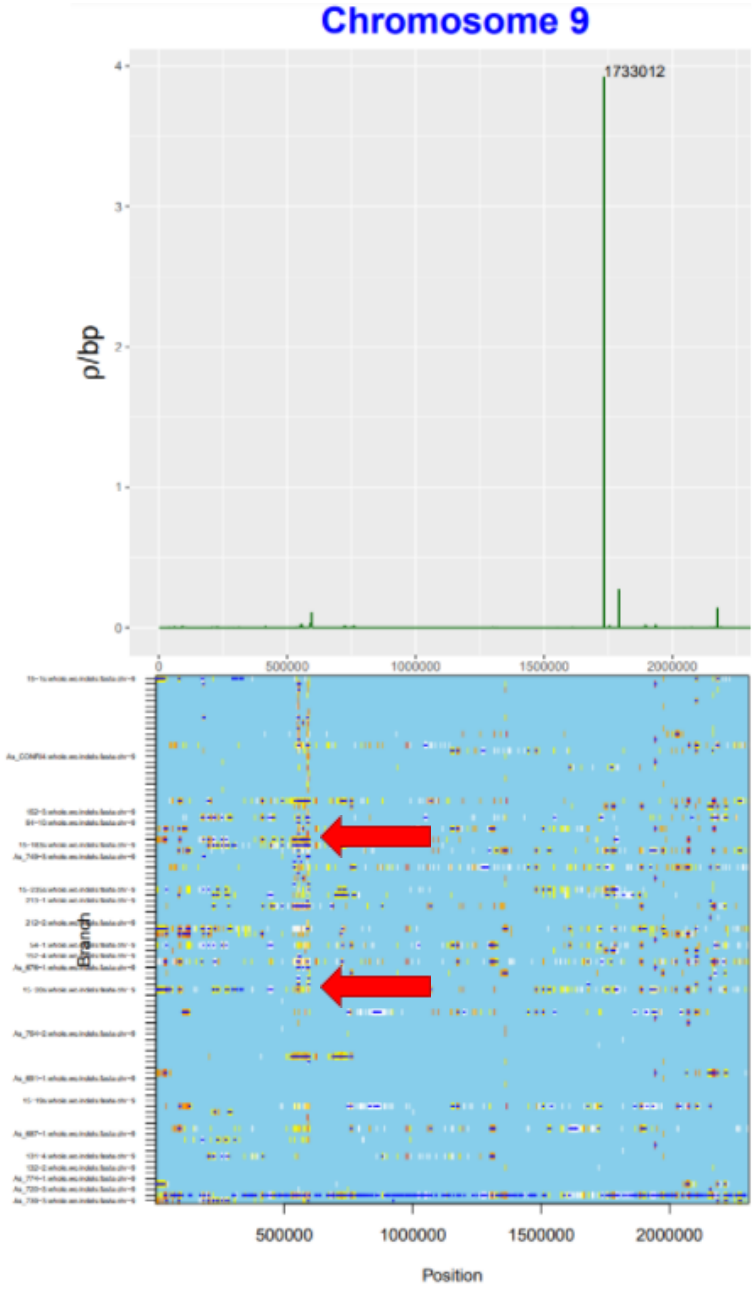


Figure 10
Chromosome 9 comparative clonalframeML LDHelmet
Top panel depicts output of LDHelmet. Lower panel shows results from clonalframeML as described in legend of figure 2.
Interestingly here the region most likely to contain recombination as predicted by clonalFrameML is hardly visible in the results by LDHelmet.

Figure 11 is interesting as the results from LDHelmet seem to point to some recombination occurring at ~500,000 mark which is corroborated by the results from clonalFrameML. The results do not have very high (ρ/bp) values and thus is not as clear a signal as seen in other chromosomes. The overall weakest evidence for recombination by LDHelmet occurs on Chromosome 5 as evidenced in **Figure 12**. The predicted locations have rather low (ρ/bp) values compared to other chromosomes but the overall locations do match up with the previous prediction by clonalFrameML.

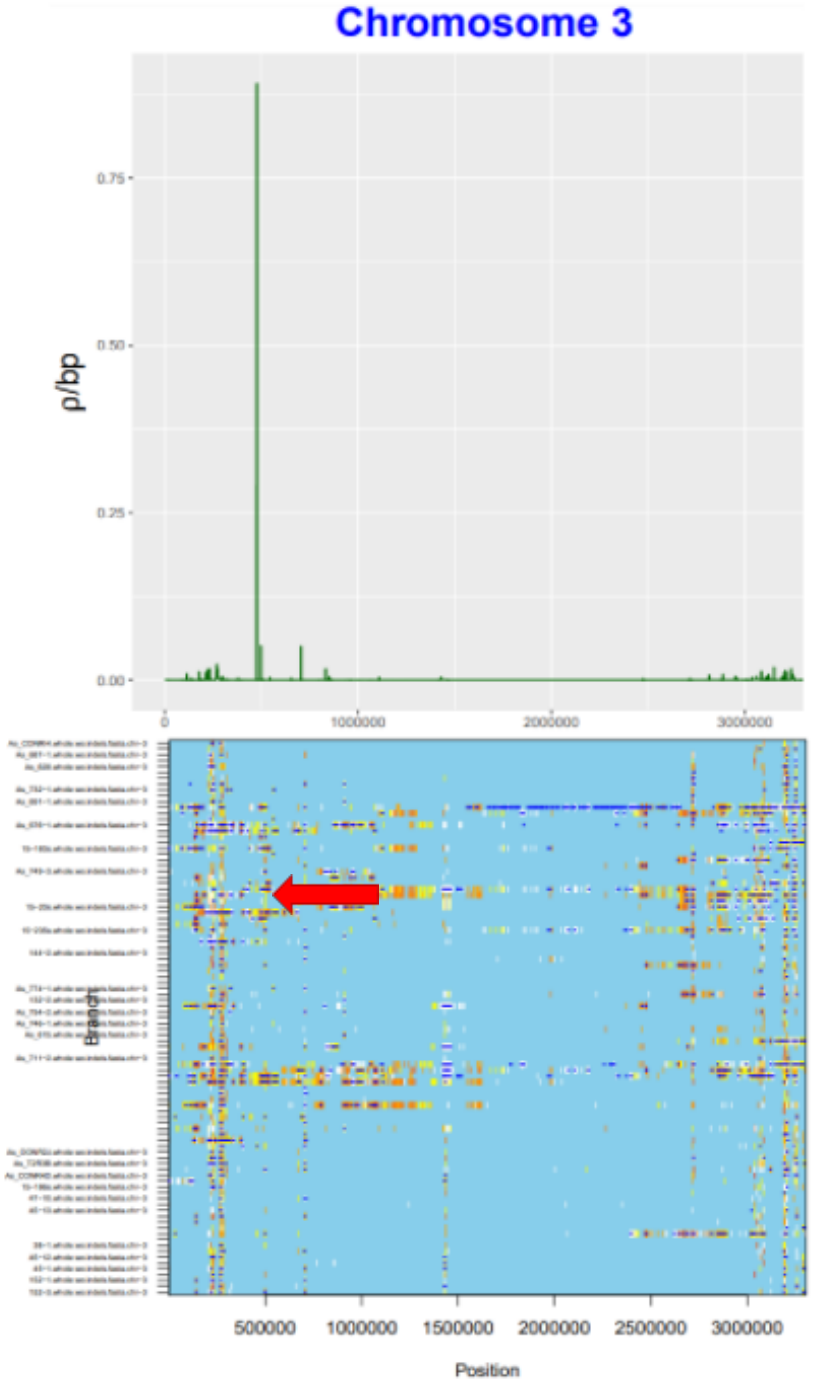


Figure 11
Chromosome 3 comparative clonalframeML LDHelmet
Top panel depicts output of LDHelmet. Lower panel shows results from clonalframeML as described in legend of figure 2.

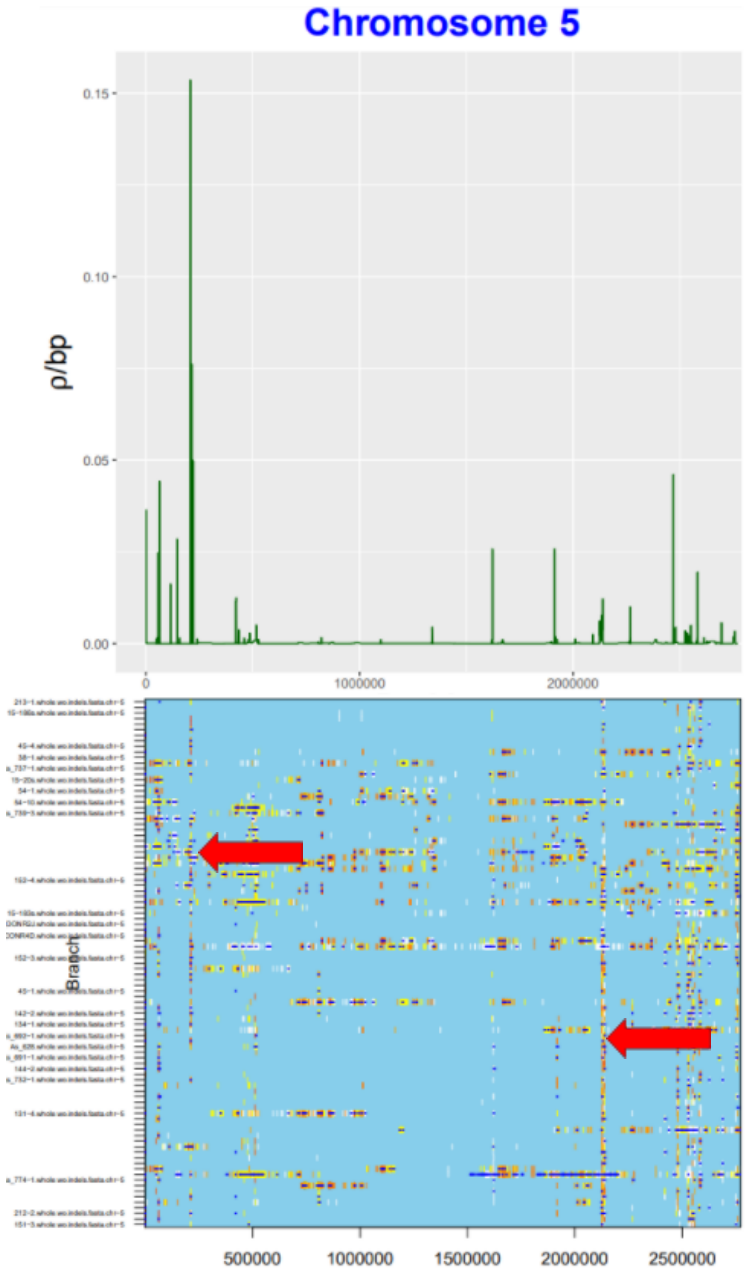


Figure 12
Chromosome 5 comparative clonalframeML LDHelmet
Top panel depicts output of LDHelmet. Lower panel shows results from clonalframeML as described in legend of figure 2.

In **Figure 13** one can see both predictions for chromosome 7 line up quite well. The predicted (ρ/bp) given by LDHelmet is rather low, however both programs do predict the occurrence of recombination in the same areas. Also of note is the overlap of recombination predictions by clonalFrameML and LDHelmet at areas ~240,000. The other predictions by LDHelmet did not have high (ρ/bp) values for the recombination events at the edges of the chromosomes as they were predicted by clonalFrameML.

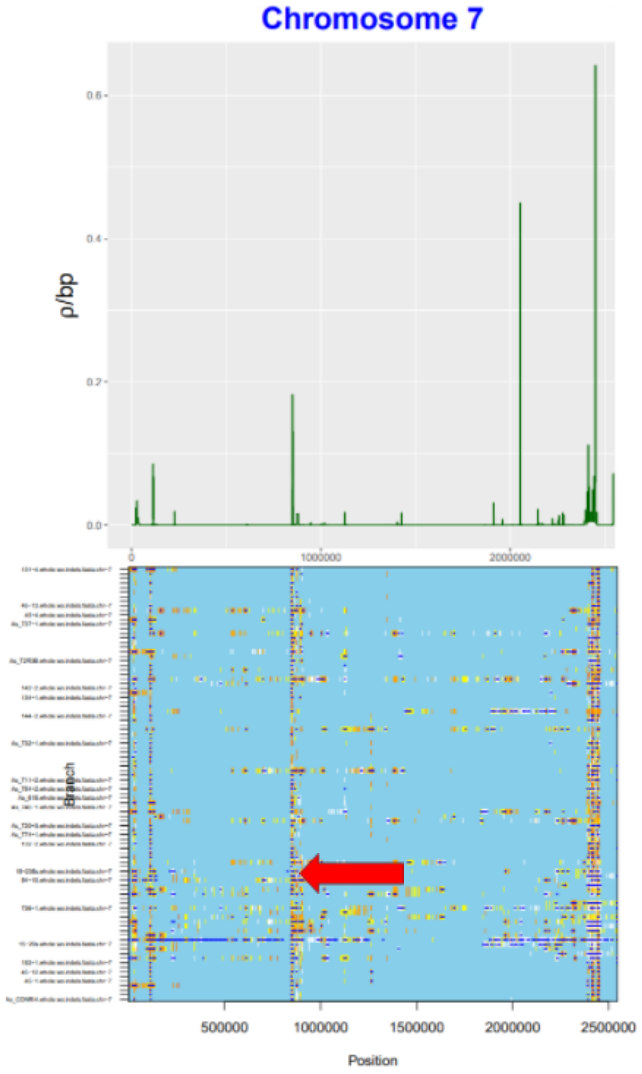


Figure 13
Chromosome 7 comparative clonalframeML LDHelmet
Top panel depicts output of LDHelmet. Lower panel shows results from clonalframeML as described in legend of figure 2.
The strong signal in the center and the very strong signal toward the edge of the chromosome are interesting. In most other chromosomes LDHelmet didn't give as high a prediction to the edge predictions often made by clonalFrameML.

left_snp	right_snp	mean	Grouped regions
931708	932838	1.6124e+00	
1336219	1336220	2.3050e+00	x
1336220	1336270	2.4553e+00	x
9823778	9824592	2.0084e+00	
10791929	10792165	1.3030e+00	
15471309	15471986	2.4595e+00	
15732088	15733962	1.5249e+00	
17601907	17603003	1.7469e+00	
18024396	18026850	1.4601e+00	
21775111	21775496	3.4462e+00	
25589391	25589859	2.9626e+00	x
25589859	25590076	2.9626e+00	x
25590076	25590320	2.9626e+00	x
25590320	25590507	2.9626e+00	x
27892523	27893641	1.8859e+00	
30324698	30325010	3.3653e+00	
32008681	32009267	2.4179e+00	

Table 2
filtered positions with mean > 1
This table holds the positions of the complete genome with a mean (ρ/bp) over 1 generated by ldhelmet.

The difficulty in interpreting these results is determining what cutoff to use as significant enough evidence. In this paper we use a cutoff value of 1 to filter out a list of relevant results from LDHelmet. Most mean (ρ/bp) were around the 10^{-5} to 10^{-6} range and a 5-fold increase in order of magnitude is unlikely to occur by chance. **Table 2** shows the SNPs across all 10 chromosomes with a mean (ρ/bp) greater than 1.
There are several SNPs that are quite small and located nearby others that are also predicted as recombination events. One SNP at 1336219 is only a single bp but is likely part of the detected recombination from 1336220 to 1336270.
With LDHelmet we find 14 regions that show evidence of recombination. Many of which upon visual inspection line up with results from clonalFrameML.

Discussion 4

Assumptions 4.1

In this paper we search for evidence of recombination occurring in *A. solani* using two tools. Several assumptions were made in order to use LDHelmet.
For LDHelmet that assumptions are: That the DNA sequences are randomly drawn from a single population, that the population follows neutral evolution, has a constant size, and a constant recombination rate across the sequence.
We believe these assumptions are met because the DNA sequences used were randomly drawn from given geographical clusters. The population fluctuates but can be said to remain relatively constant and follow neutral evolution.The recombination rate is not constant across the sequence and some chromosomes show more evidence for recombination than others but variation is expected.

Parameters 4.2

For RAxML there are several input parameters in addition to the sequence data. The -p and -x flags are random seeds for the Parsimony and Bootstrap components. The combined flag (-f a) conducts a rapid Bootstrap analysis and search for the best-scoring ML tree in one single program run. -# specifies the number of individual alternate trees used. -s is the input file location. -n is the output name. -m means that individual alpha-shape parameters, GTR rates, and empirical base frequencies will be estimated and optimized for each partition
RAxML parameters
raxmlHPC -p 1122590 -f a -x 1122590 -# 100 -s C:\Users\alex\Desktop\denbiData\bychromosome\chromosome1.fasta -n chromosome1.output -m GTRGAMMA

The input for clonalFrameML is quite simple. Its ClonalFrameML followed by the NEWICK and FASTA files as input. There are several useful flags such as -show_progress and several others such as -kappa to set the kappa (transition/transversion) value.
clonalFrameML parameters
ClonalFrameML RAxML_bipartitions.new.results all.new.exclusion.whole.wo.indels.fasta example.fasta

For LDHelmet there are numerous parameters to consider. --num_thread sets the thread count. -w sets the window size used by the **find_confs** and **pade** algorithms. The -o flag is to name the output file. -c, -p, -l are all other output file names for each step of the program. -t is the population scaled mutation rate. -r is a grid of p values (population scaled recombination rate). -x is the number of pade coefficients. -b is the block penalty for the rjmcmc algorithm and the higher the value the more likely the map tends to less variation (less sensitive to smaller effects). -n is the number of cycles that the markov chain is performed. -burn_in is the number of burn in cycles performed in addition to -n.
For the output the parameters are as follows : -m output mean, -p* outputs a variable percentile value.

LDHelmet parameters
ldhelmet find_confs --num_threads 50 -w 50 -o output.conf chromosome1.fasta
ldhelmet table_gen --num_threads 50 -t 0.01 -r 0.0 0.1 10.0 1.0 100.0 -c output.conf-o output.lk
ldhelmet pade --num_threads 50 -t 0.01 -x 11 -c output.conf -o output.pade
ldhelmet rjmcmc --num_threads 50 -l output.lk -p output.pade -s chromosome1.fasta -w 50 -b 50.0 --burn_in 100000 -n 1000000 -o LDHelmetoutputchr1.post
ldhelmet post_to_text -m -p 0.025 -p 0.50 -p 0.975 -o LDHelmetoutputchr1.txt LDHelmetoutputchr1.post
ldhelmet max_lk --num_threads 50 -l output.lk

Runtime Factors 4.3
ClonalFrameML was designed for bacterial genomes and intended to run in 15-30 minutes. For this dataset it ran about 4-5 hours on desktop per chromosome. The lack of parallelization was also rather frustrating and is also a logged request to the developer.
LDHelmet was run on a server provided by Stam lab with 50 CPU cores and 128GB RAM and took ~10 hours per run. Fortunately despite the vastly longer complexity, compute and RAM requirements, the multi-threading cut the time down dramatically.

Confidence in findings 4.4

The results achieved in this paper do seem to show evidence for some occurrence of recombination across the genome. Possible false positives detected at either ends of the chromosomes by clonalFrameML are also detected by LDHelemt, but with much lower (ρ/bp). Both tools have been tested on many datasets containing no recombination and do not generate results with visible peaks/blue lines. Thus the occurrence of these predicted recombination events at the ends of chromosomes seems likely to be due to the biological reality of the 3d structure of a chromosome. The overlap between the results of both these tools suggests the historical occurrence of recombination in the *Alternaria solani* genome.
While many of the results do overlap between both results, not all do. As is evidenced in the LDHelmet results from chromosome 8, there are some positions found by LDHelmet that are not evident in the clonalFrameML results and vica versa. Small differences are to be expected due to them being different tools with different methods for predicting recombination events. With different tools weighing certain factors differently this is bound to cause some differences. For the most part however a large spike seen in LDHelmet is reflected in clonalFrameML.

Possible extensions 4.5

In the future it would be interesting to take a look at the differences in recombination rates across each chromosome. Also interesting would be a closer look at the exact bp configurations at the locations where clonalFrameML, LDHelmet, or other programs overlap. This would perhaps allow for a further reduction of noise if artifacts such as repeating sequences at the ends of chromosomes were filtered out.

References

1 - **Milgroom MG, Jiménez-Gasco M del M, García CO, Drott MT, Jiménez-Díaz RM**. **2014**. Recombination between Clonal Lineages of the Asexual Fungus Verticillium dahliae Detected by Genotyping by Sequencing. *PLOS ONE* **9**: e106740.
2 - **Leiminger JH, Hausladen H**. **2012**. Early Blight Control in Potato Using Disease-Orientated Threshold Values. *Plant Disease* **96**: 124–130.
3 - **Adhikari P, Oh Y, Panthee DR**. **2017**. Current Status of Early Blight Resistance in Tomato: An Update. *International Journal of Molecular Sciences* **18**: 2019.
4 - **McGovern,Dr Robert, Hall F**. DISEASE MANAGEMENT: Early Blight of Tomato. https://ipm.ifas.ufl.edu/resources/success_stories/T&PGuide/pdfs/Chapter5/Early_Blight.pdf accessed 16/6/2021
5 - **FRAC HOME**: FUNGICIDE RESISTANCE ACTION COMMITTEE, https://www.frac.info/frac-teams/working-groups/sdhi-fungicides/information, accessed 5/7/2021
6 - **Einspanier S, Susanto T, Metz N, Wolters PJ, Vleeshouwers VGAA, Lankinen Å, Liljeroth E, Landschoot S, Ivanović Ž, Hückelhoven R, *et al.* 2021**. Whole genome sequencing elucidates the species-wide diversity and evolution of fungicide resistance in the early blight pathogen Alternaria solani.
7 - **Stam R, Einspanier S, Susanto T. 2021**. Supplementary Data for: Whole genome sequencing elucidates the species-wide diversity and evolution of fungicide resistance in the early blight pathogen Alternaria solani.
8 - **Wolters PJ, Faino L, van den Bosch TBM, Evenhuis B, Visser RGF, Seidl MF, Vleeshouwers VGAA**. **2018**. Gapless Genome Assembly of the Potato and Tomato Early Blight Pathogen Alternaria solani. *Molecular Plant-Microbe Interactions*® **31**: 692–694.
9 - **Li H, Durbin R**. **2009**. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**: 1754–1760.
10 - **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* 2010**. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.
11 - **Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. 2012a.** Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Frontiers in Genetics
12 - **Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012b**. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly
13 - **RAxML** - **A. Stamatakis 2014**. "RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies". In *Bioinformatics*, 2014, open access.
14 - **clonalFrameML** - **Didelot X, Wilson DJ**. **2015**. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Computational Biology* **11**: e1004041.
15 - **LDHelmet** - **Chan AH, Jenkins PA, Song YS**. **2012**. Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLOS Genetics* **8**: e1003090.

16 - **Wickham, H. 2016**. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org
17 - **Ape** - **ape – *ape***. **2021**. cran.
18 **- Pharagorn** - Schliep K.P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics, 27(4) 592-593

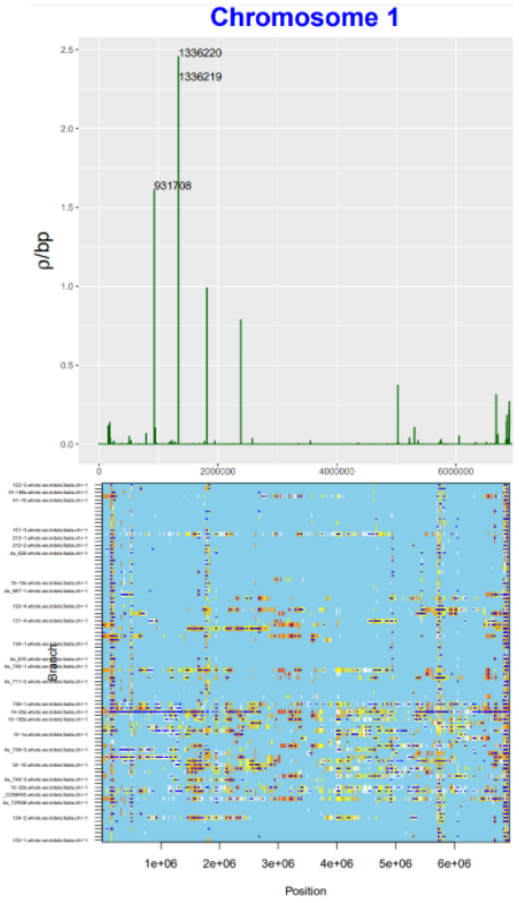
Appendix

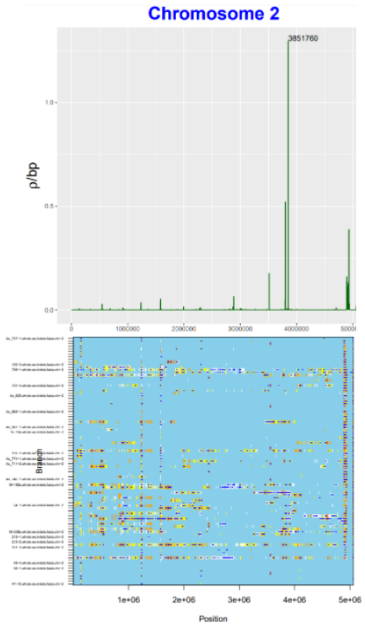
List of Figures

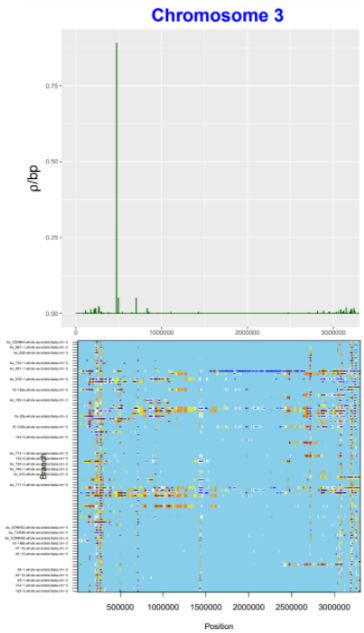
List of Tables

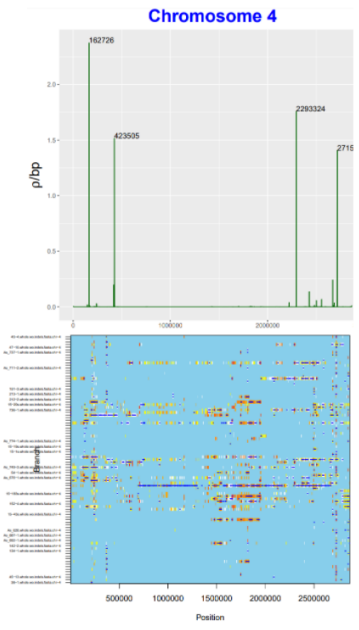
Supplementary figures

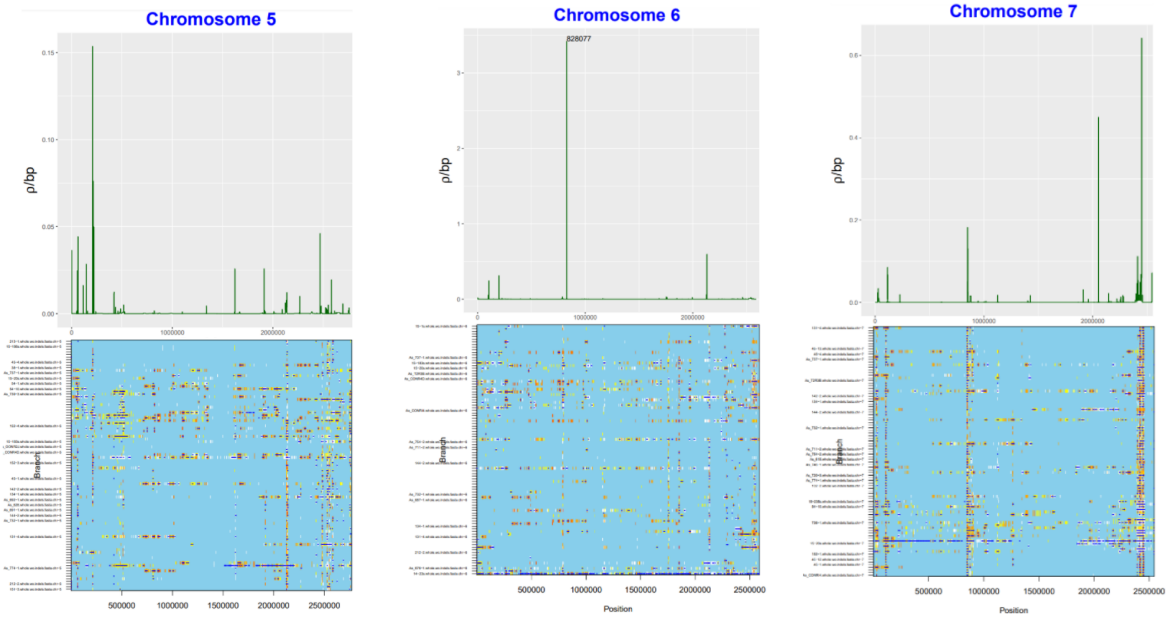
Visualized comparison clonalFrameML vs LDHelmet chromosomes 1-10

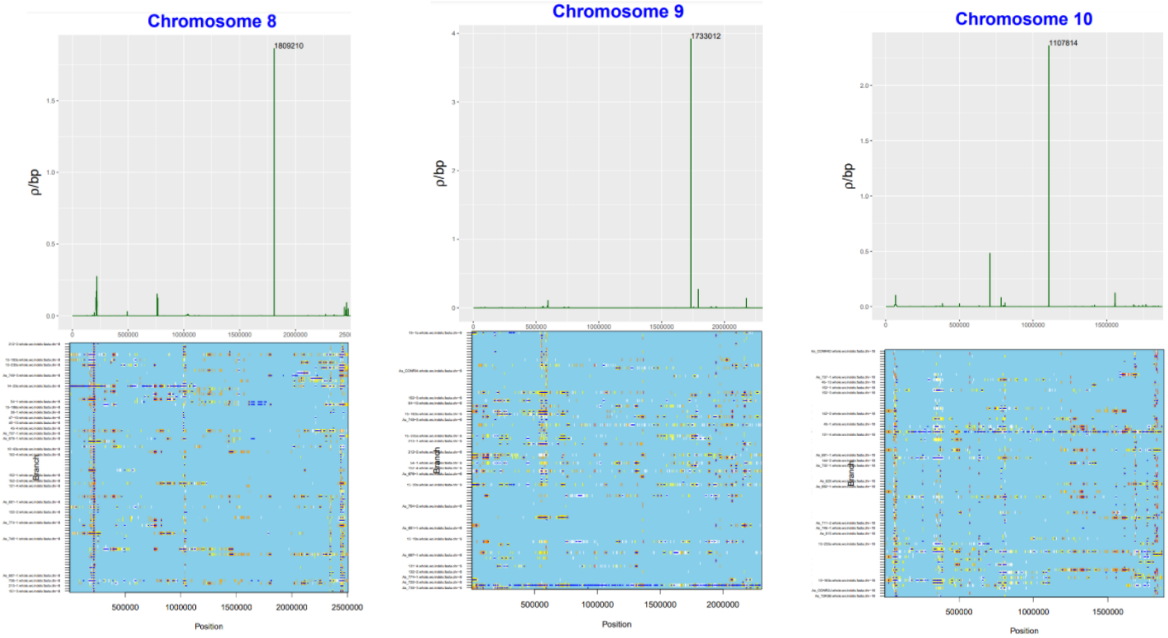

Chromosome 1
1336220
1336219
931708
ρ/bp
0
2000000
4000000
6000000
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Position


Chromosome 2
3851760
ρ/bp
1e+06
2e+06
3e+06
4e+06
5e+06
Position


Chromosome 3
ρ/bp
500000
1000000
1500000
2000000
2500000
3000000
Position


Chromosome 4
162726
423505
2293324
ρ/bp
500000
1000000
1500000
2000000
2500000
Position


Chromosome 5
Chromosome 6
Chromosome 7
ρ/bp
Position


Chromosome 8
Chromosome 9
Chromosome 10
ρ/bp
Position