To do

Neutral case: Simulations of a sweep could use a very large population. For the neutral case, I need to consider a smaller population - for example, 100 individuals for 1000 generations. This could be spliced onto a brief sweep.

Genealogies: I need to store genomes in every generation, and then reconstruct genealogies along the genome. The data needs to be savagely pruned, however.

SNP: Given the genealogies, throw down SNP onto the genomes.

Data formats: I need the current junctions scheme; tree-sequence; and Newick.

It is rather easy to identify branches, as ancestral to particular sets of genomes. These have some persistence along the genome, and through time.

Reconstructing genealogies is non-trivial, but might be best done by identifying branches. There could be an extremely large number of trivially different genealogies..

One can also throw down SNP relatively easily, just by setting a mutation rate. This can also be done by branches, as long as one ensures that the set of all branches fills the whole ARG.

Example: s=0.025, R=0.05, $2N = 10^6$, 100 genomes

Neutral case

Setting up simulations

Iterating back for 10×2N generations; n=2N=100; R=0.05

Iterating back until coalescence; n=2N=100; R=0.1

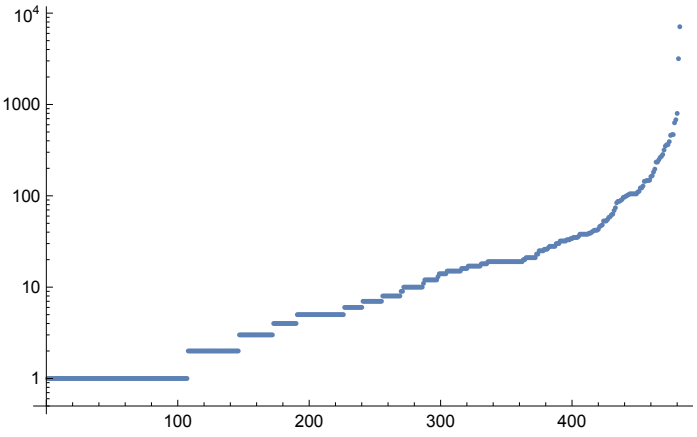
Identifying branches

Identifying branches

Here, each instance is a lineage and a generation. The extent of a branch along the genome and through time may be complex. The numbers of times each branch appears is highly skewed. Two branches appear >1000 times, but there are 85 instances once, and the median is 7.

In[407]:=
```
tb = branchList[pl2];
ListLogPlot[Sort[Last /@ tb], PlotRange → All]
{Length[pl2], Length[tb],
  Total[Last /@ tb], Median[Last /@ tb], Mean[Last /@ tb]} // N
```

Out[408]=

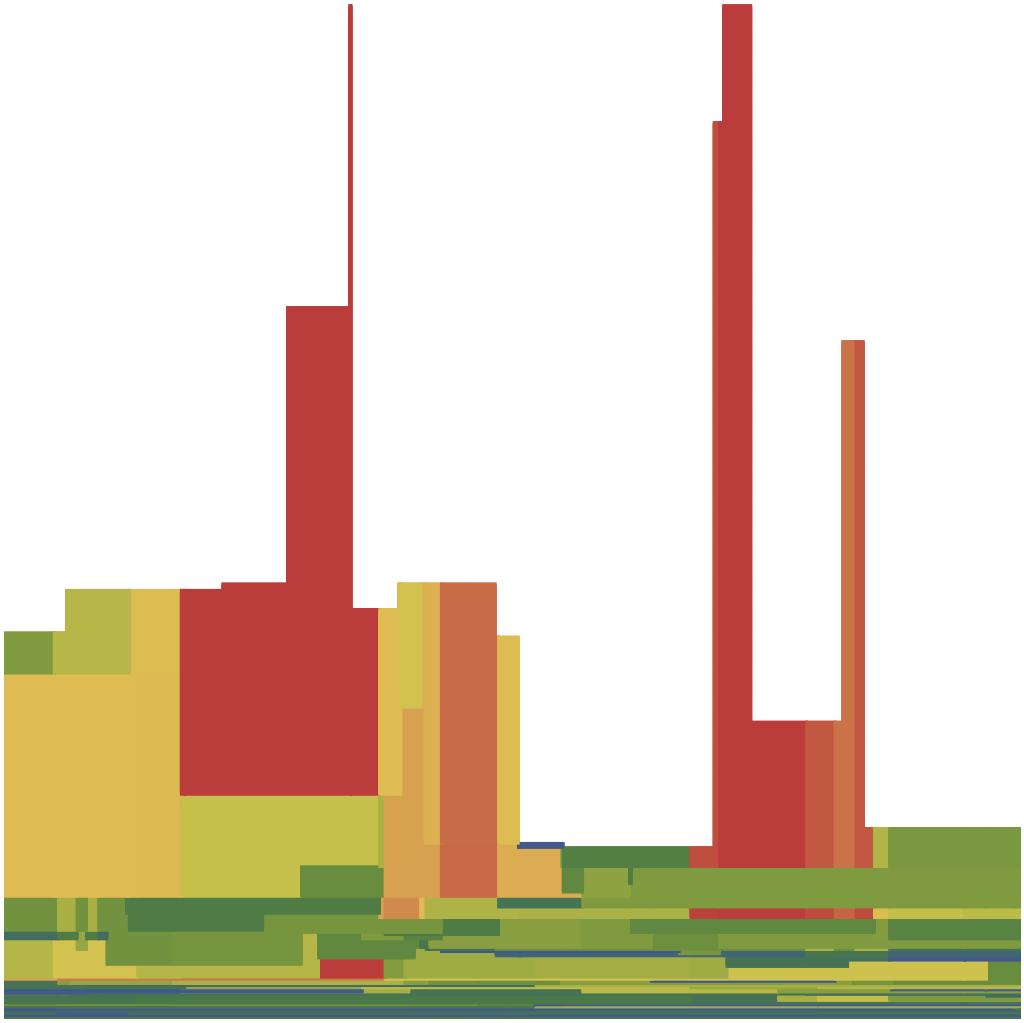


Out[409]= {478., 482., 25 439., 7., 52.778}

This plots 478 generations on the vertical axis, and ma length $R = 0.1$ on the horizontal. of the 482 blocks is coloured in according to the number of instances (red is high)(approximately the # of generations). However, many are overlaid in this representation. The white areas have coalesced.

In[1343]:=
```
tb = branchList[pl2];
mx = Max[Log[Last /@ tb]]; cf = ColorData["DarkRainbow"];
Show[plotBranch[#[[1]], cf[Log[#[[2]]]/mx], pl2, 0.1] & /@ tb,
 PlotRange → {{0, 0.1}, {0, 478}}, AspectRatio → 1]
```

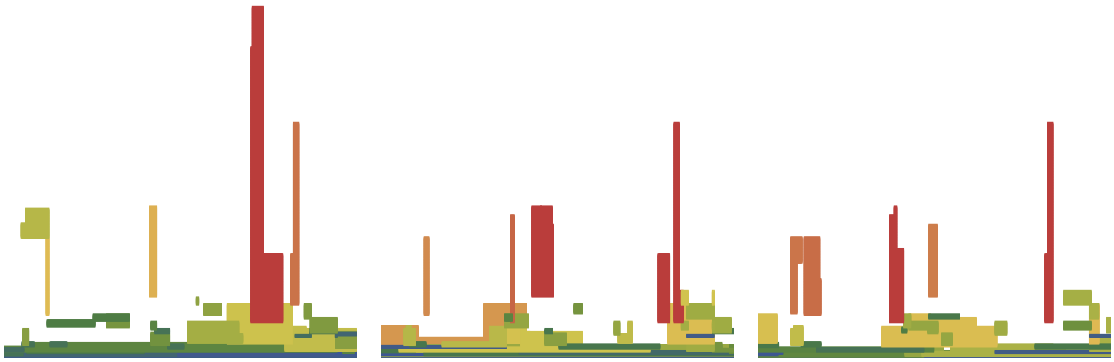
Out[1345]=



This shows three sets of 1 in 10 lineages. Again, colour corresponds to timespan. Note the inverse relation between block length and timespan:

In[480]:=
```
tbs[j_, k_] := branchList[pl2][[j ;; -1 ;; k]];
cf = ColorData["DarkRainbow"];
gr[j_, k_] := Module[{mx = Max[Log[Last /@ tbs[j, k]]]},
  Show[plotBranch[#[[1]], cf[Log[#[[2]]]/mx], pl2, 0.1] & /@ tbs[j, k],
   PlotRange → {{0, 0.1}, {0, 478}}, AspectRatio → 1]];
GraphicsRow[gr[#, 10] & /@ {1, 2, 3}]
```

Out[482]=



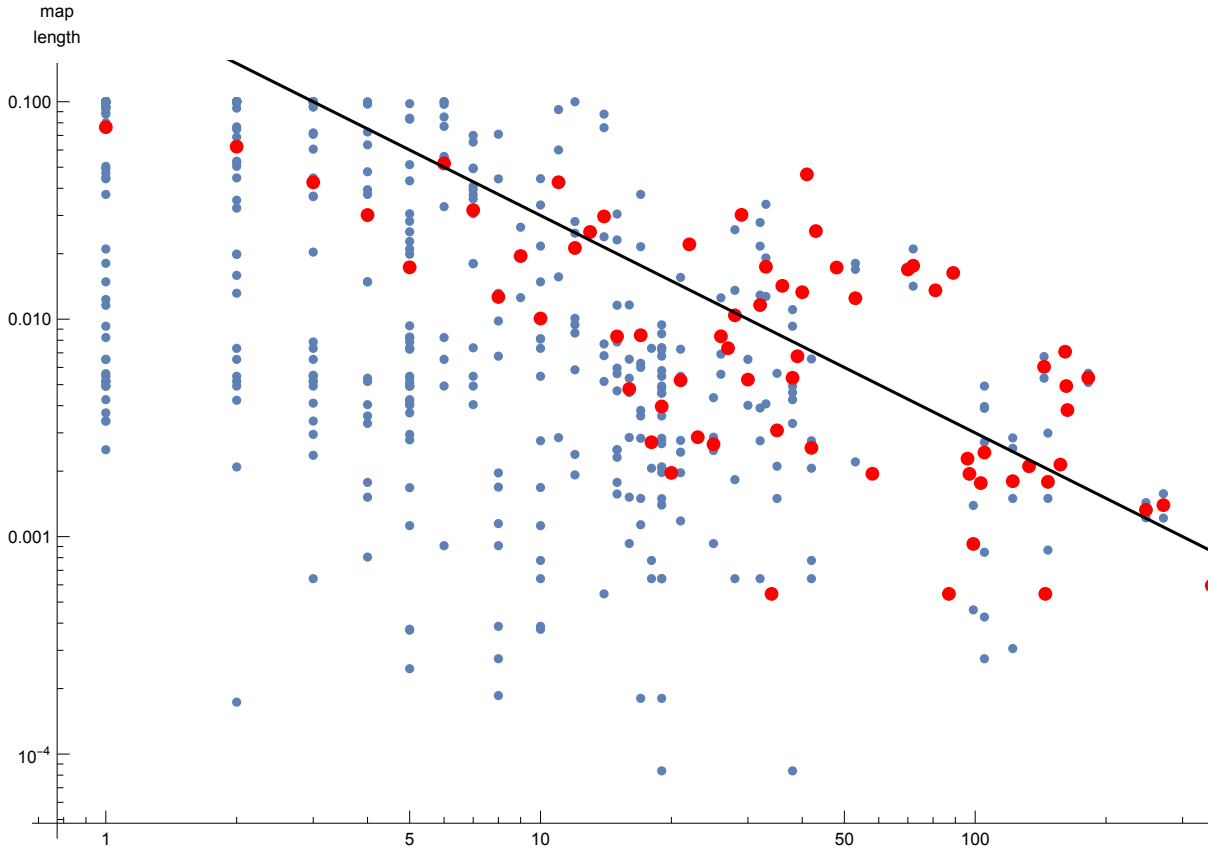
inverse relation between block length and timespan

This shows the mean block length against its timespan. The 482 branches are shown by blue dots, and the mean for each timespan is shown by the red dots. There is an inverse relation, but as ~ $t^{-0.6}$, rather than the expected t^{-1}. The two fits are to the blue vs the red dots; the black line is 0.3 t^{-1}, to indicate the slope expected with a simple inverse relation.

```
txl = (pb = posBlock[pl2, #[[1]], 0.1];
     {Length[Union[pb[[All, 1]]], Mean[pb[[All, 2]].{-1, 1}]}) & /@ tb;
txlm = Mean /@ GatherBy[txl, Log[#[[1]]] &];
```

In[727]:=
```
Show[ListLogLogPlot[txl], ListLogLogPlot[txlm, PlotStyle → Red],
 LogLogPlot[0.3 t^-1, {t, 1, 500}, PlotStyle → Black],
 AxesLabel → {"time\nspan", "map\nlength"}]
{Fit[Log[txl], {1, t}, t], Fit[Log[txlm], {1, t}, t]}
```

Out[727]=

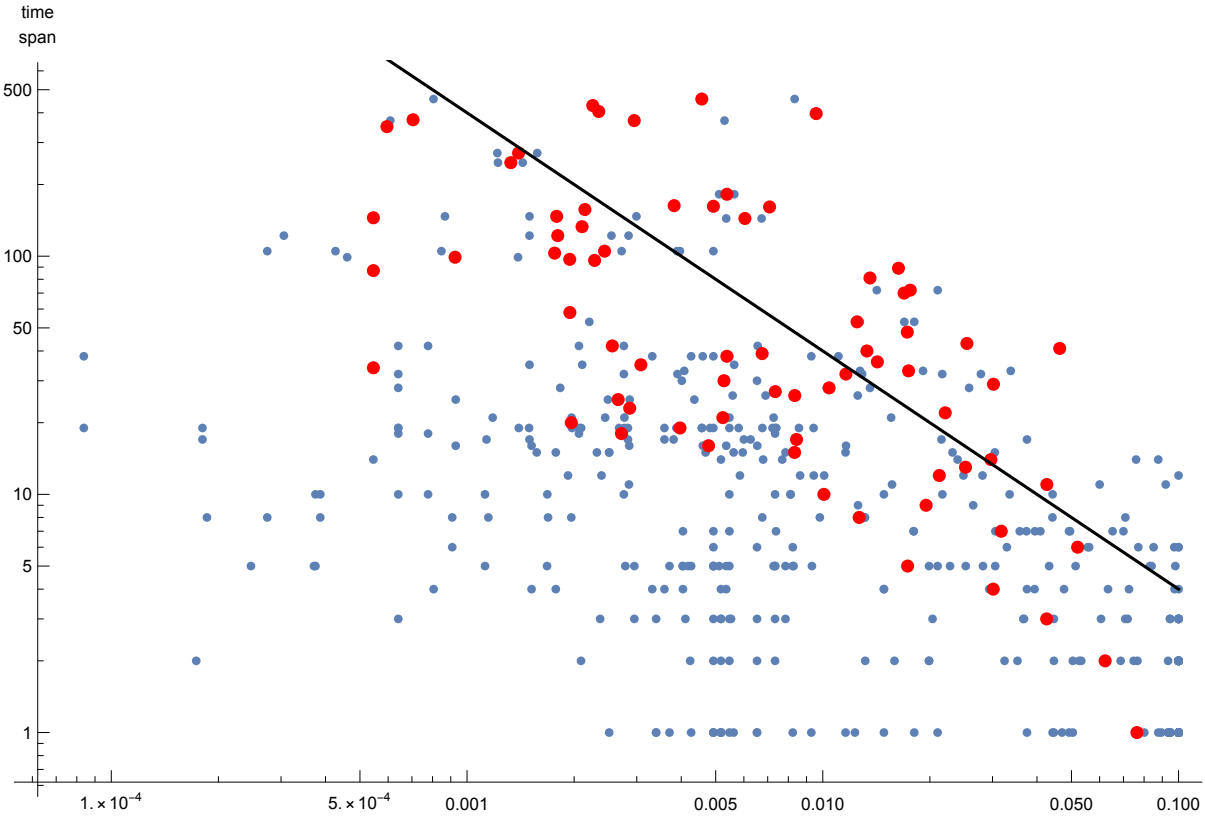


Out[728]= $\{-3.1815 - 0.657731\, t, -2.76384 - 0.609253\, t\}$

This shows timespan plotted against map length - it is not obvious which direction causality runs. Now, we fit a relation $t \sim r^{-0.7}$ which contradicts the above fit, $r \sim t^{-0.6}$. Maybe, a direct inverse relation $t \sim r^{-1}$ is possible ? The black line is $0.4\, r^{-1}$, to indicate the slope expected with a simple inverse relation

In[735]:=
```
Show[ListLogLogPlot[Reverse /@ txl],
 ListLogLogPlot[Reverse /@ txlm, PlotStyle → Red],
 LogLogPlot[0.4 x^-1, {x, 10^-4, 0.1}, PlotStyle → Black],
 AxesLabel → {"map\nlength", "time\nspan"}]
{Fit[Log[Reverse /@ txl], {1, r}, r], Fit[Log[Reverse /@ txlm], {1, r}, r]}
```

Out[735]=



Out[736]= $\{-0.475329 - 0.547026\, r,\ 0.138248 - 0.717338\, r\}$

email to Miso

Throwing down SNP for μ/r=10 (mostly)

Throwing down SNP: μ/r=100

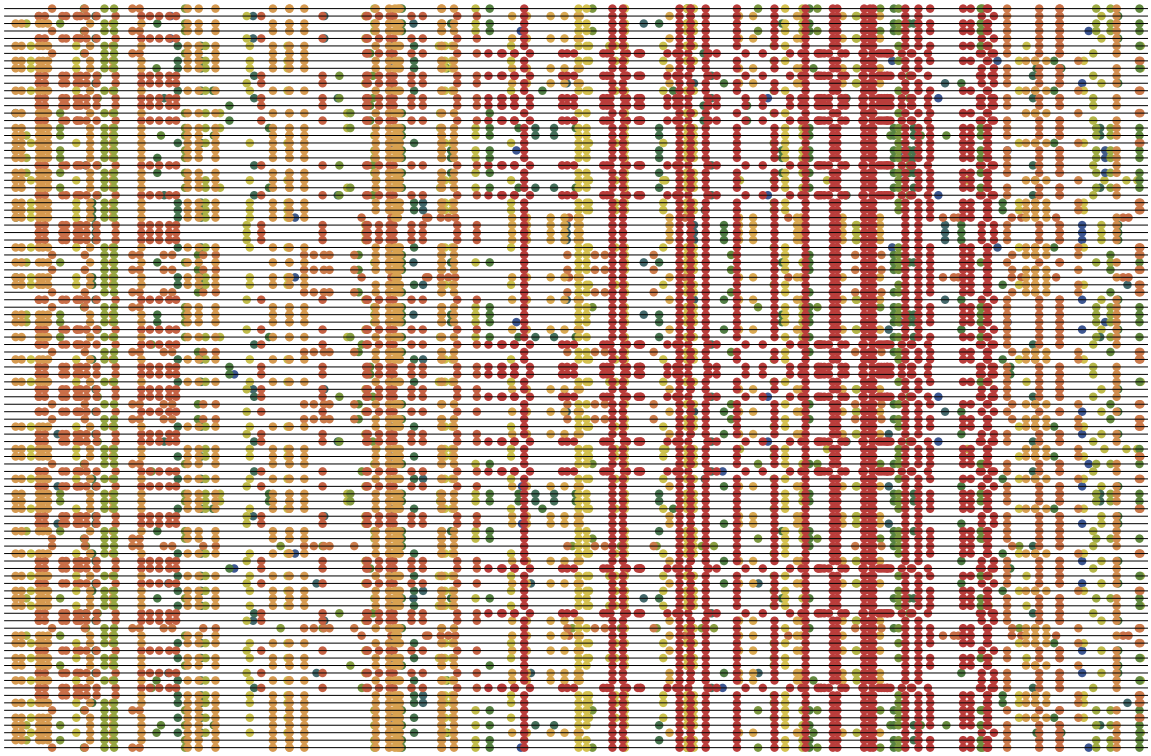
Throwing down SNP: μ/r=10

Plotting all the SNP

This plots the full set of SNP, with colours indicating the time spanned by the branch (red is long, blue is short). Long branches tend to span a shorter distance along the genome, but nevertheless, most SNP are on long branches. Tips and the common ancestor are not shown.

In[1135]:=
```
bl = branchLength[posBlock[pl2, #[[1]], 0.1]] & /@ tb;
cc = cf /@ Log[bl] / Max[Log[bl]];
gg =
 Graphics[Join[plotSNP[cc, 0.003, pop], plotGenomes[100, 0.1]], AspectRatio → 0.3]
```

Out[1136]=



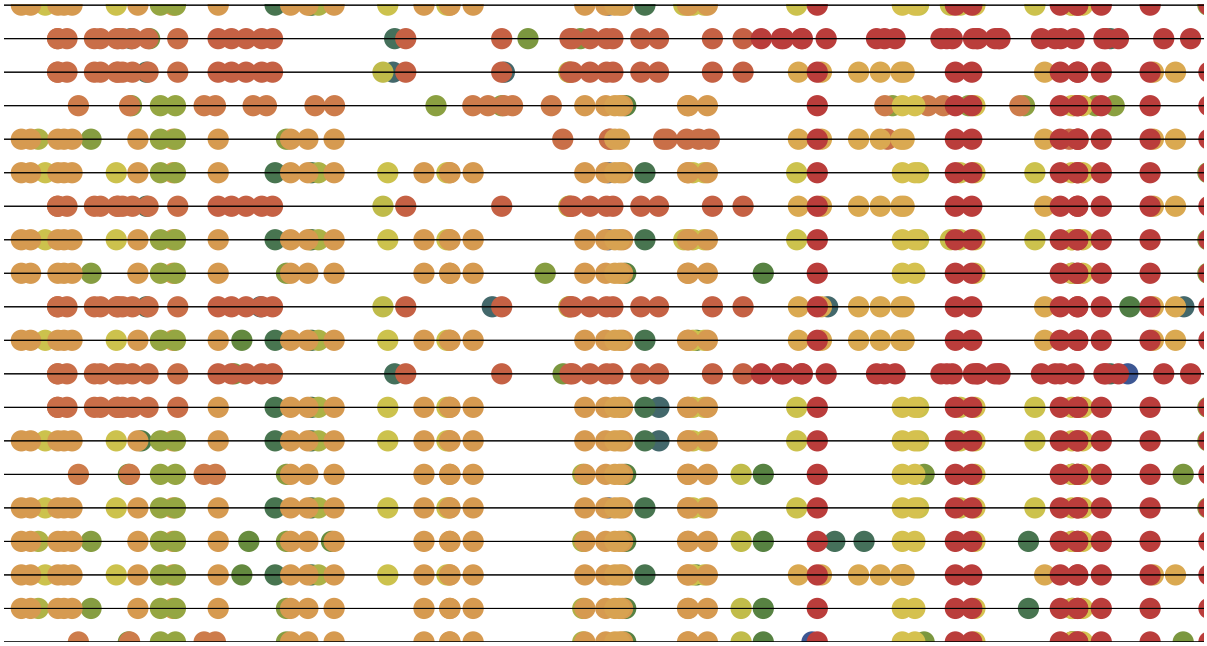
This shows an inset of 20 genomes on {0,0.05}. Most SNP are on long branches (red)

In[1109]:=
```
ColorData["DarkRainbow", "Image"]
```

Out[1109]=

In[1139]:=
```
Show[Graphics[Join[plotSNP[cc, 0.01, pop], plotGenomes[100, 0.1]],
  AspectRatio → 0.3], PlotRange → {{0, 0.05}, {1, 20}}]
```

Out[1139]=



Finding the 24 branches with ≥10 SNP

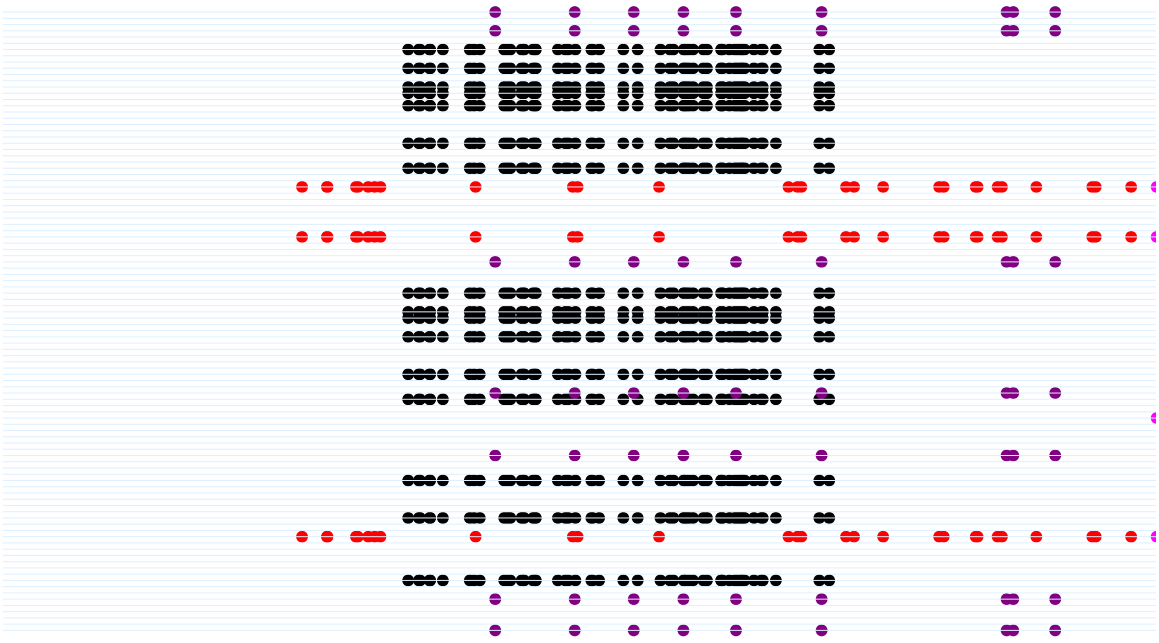
Plotting the SNP on the most substantial branches

These plots show that there are “interesting” regions. Are these just those regions with deepest genealogies and tehrefore the highest density of NP?

These are six of the most substantial SNP:

In[1175]:=
```
ggBig = Graphics[Join[{PointSize[0.005]},
    Join @@ MapThread[{#2, plotSNP[#1, pop]} &,
      {{218, 55, 311, 359, 56, 65}, {Black, Red, Green, Cyan, Magenta, Purple}}],
    {LightBlue}, plotGenomes[100, 0.1]], AspectRatio → 0.3]
```

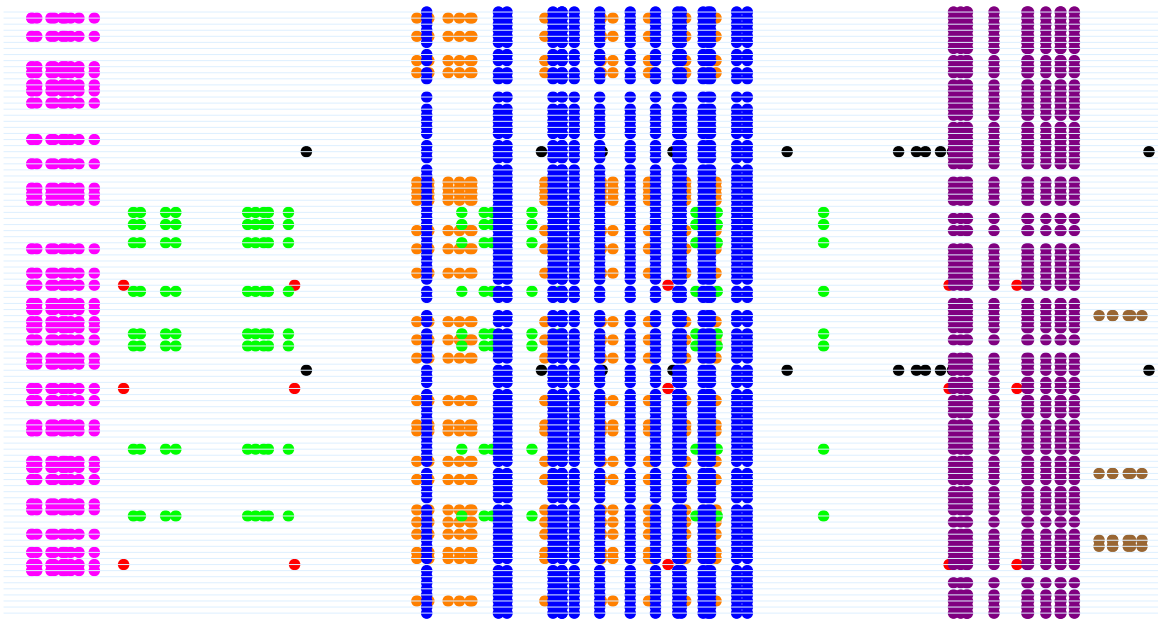
Out[1175]=



These are the remaining nine:

In[1178]:=
```
ggBig = Graphics[Join[{PointSize[0.005]},
    Join @@
     MapThread[{#2, plotSNP[#1, pop]} &, {{26, 48, 52, 190, 209, 211, 214, 371, 375},
       {Black, Brown, Red, Orange, Green, Cyan, Magenta, Purple, Blue}}],
    {LightBlue}, plotGenomes[100, 0.1]], AspectRatio → 0.3]
```

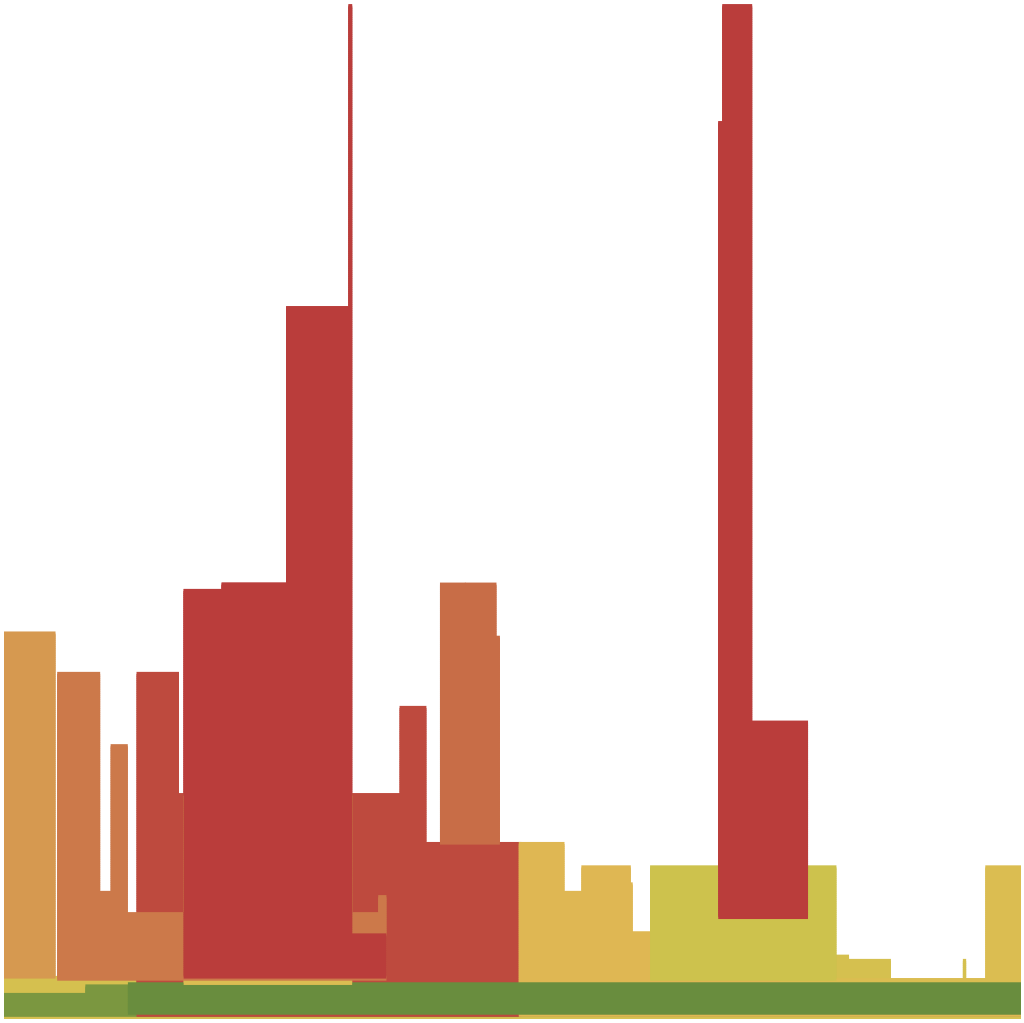
Out[1178]=



This is the block structure, plotting map length (x axis) against time (y axis). White areas have coalesced.

In[1346]:=
```
tb = branchList[pl2];
mx = Max[Log[Last /@ tb]]; cf = ColorData["DarkRainbow"];
Show[plotBranch[#[[1]], cf[Log[#[[2]]]/mx], pl2, 0.1] & /@ tb[[big]],
 PlotRange → {{0, 0.1}, {0, 478}}, AspectRatio → 1]
```

Out[1348]=



Superimposing SNP onto blocks

This shows the 53 SNP on branch #218, superimposed on the haplotype block:

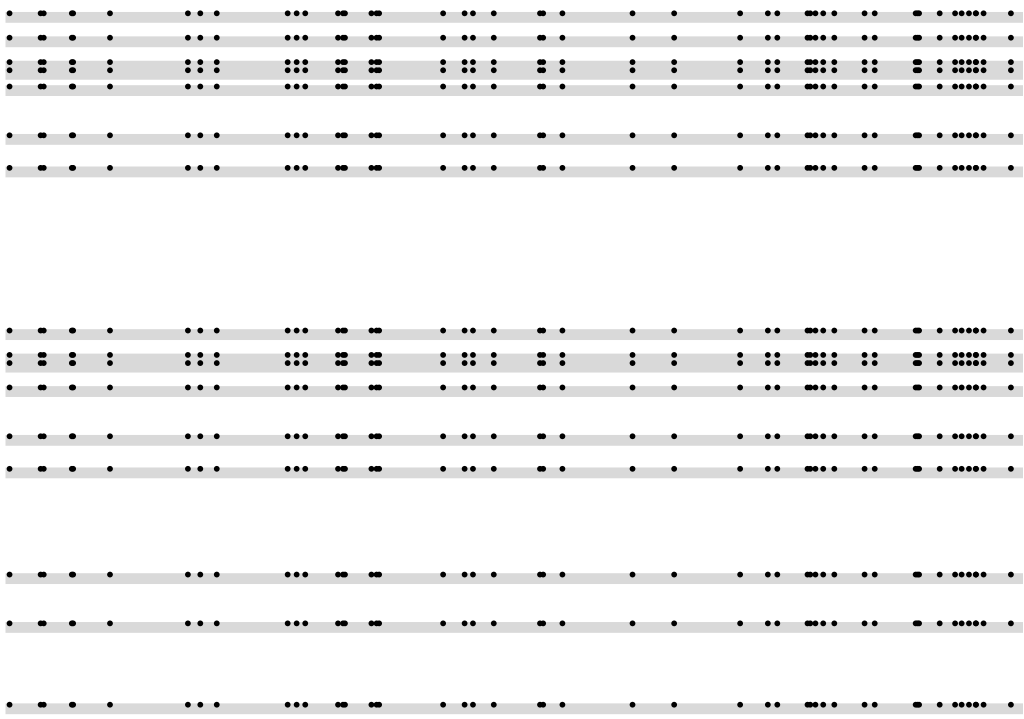
In[1317]:=
```
Show[plotBlock[tb[[218, 1]], LightGray, pl2, 0.1],
 Graphics[{Black, plotSNP[218, pop]}],
 AspectRatio → 1, PlotRange → {{0, 0.1}, {0, 101}}]
```

Out[1317]=

This shows the same, but looking at the region {0.015,0.04}

In[1322]:=
```
Show[plotBlock[tb[[218, 1]], LightGray, pl2, 0.1],
 Graphics[{Black, plotSNP[218, pop]}],
 AspectRatio → 0.5, PlotRange → {{0.015, 0.04}, {0, 101}}]
```

Out[1322]=



Throwing down SNP for μ/r=2

Throwing down SNP: μ/r=2

Plotting all the SNP

This plots the full set of SNP, with colours indicating the time spanned by the branch (red is long, blue is short). Long branches tend to span a shorter distance along the genome, but nevertheless, most SNP are on long branches. Tips and the common ancestor are not shown.

In[1327]:=
```
bl = branchLength[posBlock[pl2, #[[1]], 0.1]] &/@ tb;
cc = cf /@ Log[bl] / Max[Log[bl]];
gg = Graphics[
  Join[plotSNP[cc, 0.003, pop2], plotGenomes[100, 0.1]], AspectRatio → 0.3]
```

Out[1328]=

This is the colour key: red indicates the longest branches.

In[ ]:=
```
ColorData["DarkRainbow", "Image"]
```

Out[ ]=

Finding the 24 branches with ≥10 SNP

There are 148 SNP on 380 branches, of which 32 are on the longest 7 branches (which have ≥4 SNP):

In[1339]:=
```
big2 = Pick[Range[Length[tb]], (Length[#] ≥ 4) &/@ snp2];
{{Length[tb], Length[big2]},
  {Total[Length /@ snp2], Total[Length /@ snp2[[big2]]]}} // TableForm
```

Out[1340]//TableForm=

380	7
148	32

These are the index; the # SNP (up to 6); the timespan. There is an inverse relation between map length and timespan for these 7 longest branches

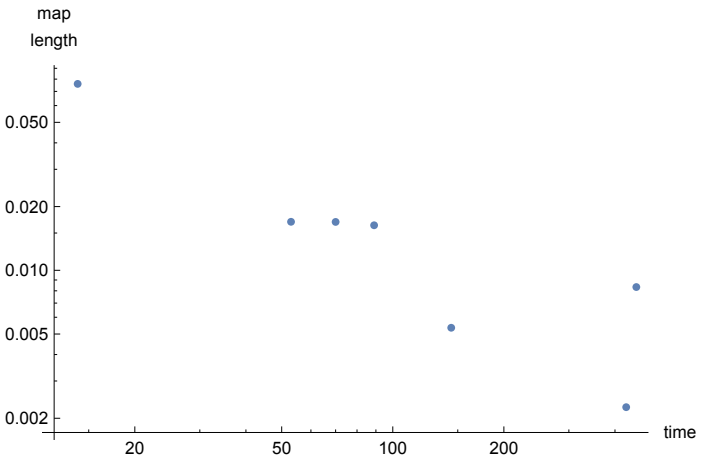
In[1335]:=
```
tt = {big2, Length /@ snp2[[big2]], branchLength[posBlock[pl2, #[[1]], 0.1]] & /@
    tb[[big2]], branchWidth[posBlock[pl2, #[[1]], 0.1]] & /@ tb[[big2]]};
tt //
 TableForm
```

Out[1336]//TableForm=

48	52	190	208	210	218	311
5	6	4	4	5	4	4
70	14	89	144	53	457	429
0.0169243	0.075972	0.0163205	0.00535247	0.0169481	0.00833454	0.0(

In[1337]:=
```
ListLogLogPlot[tt[[{3, 4}]] // Transpose,
 PlotRange → All, AxesLabel → {"time", "map\nlength"}]
Fit[Log[tt[[{3, 4}]] // Transpose], {1, t}, t]
```

Out[1337]=



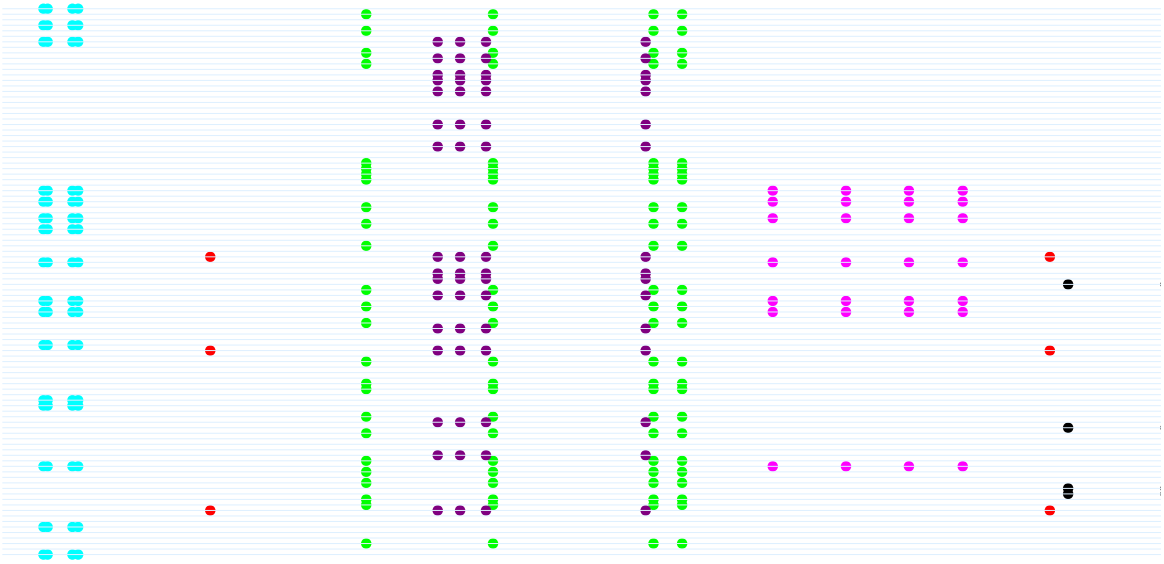
Out[1338]= $-0.647086 - 0.81321\ t$

Plotting the SNP on the most substantial branches

These plots show that there are “interesting” regions. Are these just those regions with deepest genealogies and therefore the highest density of NP?

In[1341]:=
```
ggBig = Graphics[Join[{PointSize[0.005]},
    Join @@ MapThread[{#2, plotSNP[#1, pop2]} &,
      {big2, {Black, Red, Green, Cyan, Magenta, Purple, Blue}}],
    {LightBlue}, plotGenomes[100, 0.1]], AspectRatio → 0.3]
```

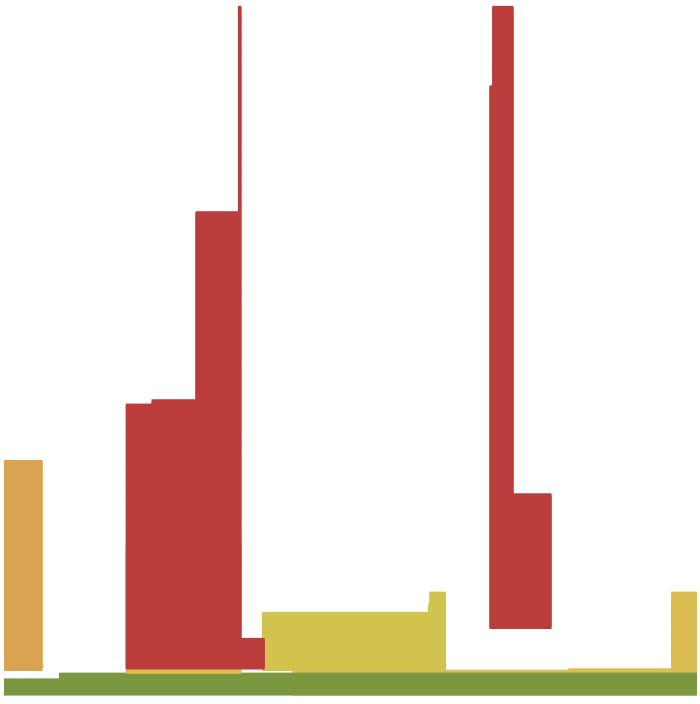
Out[1341]=



This is the block structure for these 7 top SNP, plotting map length (x axis) against time (y axis). White areas have coalesced.

In[1349]:=
```
mx = Max[Log[Last /@ tb]]; cf = ColorData["DarkRainbow"];
Show[plotBranch[#[[1]], cf[Log[#[[2]]] / mx], pl2, 0.1] & /@ tb[[big2]],
 PlotRange → {{0, 0.1}, {0, 478}}, AspectRatio → 1]
```

Out[1350]=



Superimposing SNP onto blocks

This shows the 6 SNP on branch #52, superimposed on the haplotype block. This branch spans almost the whole

In[1342]:=
```
Show[plotBlock[tb[[52, 1]], LightGray, pl2, 0.1],
 Graphics[{Black, plotSNP[52, pop]}],
 AspectRatio → 1, PlotRange → {{0, 0.1}, {0, 101}}]
```

Out[1342]=

This is the block structure of block 52, plotting map length (x axis) against time (y axis).

In[1353]:=
```
mx = Max[Log[Last /@ tb]]; cf = ColorData["DarkRainbow"];
Show[plotBranch[tb[[53, 1]], cf[Log[tb[[53, 2]]] / mx], pl2, 0.1],
 PlotRange → {{0, 0.1}, {0, 478}}, AspectRatio → 1]
```

Out[1354]=

Regression to the mean

Definitions

- probSurv
- probCoal
- makeReps
- makeRepsFix
- makeF
- iterF
- findT
- h[ρ]
- soln
- solnApprox

solnTr

coalesce

recombine

iterBack

makeFamilies

iterC

makeC

deleteAllNC

deleteNC

deleteNA

condense

randomSet

coalesceC

splitAncestry

recombineC

pos

findBlocks

findAllBlocks

plotAncestry

coalescedQ

int

branchList

posBlock

plotBranch, plotBlock

makeSNP

addSNP

branchLength, branchWidth

plotSNP, plotGenomes

Regression to the mean