Accessing invisible information in a large scale heatmap

Navigating in LD heatmap for thousands of SNPs

Jieun Jeong

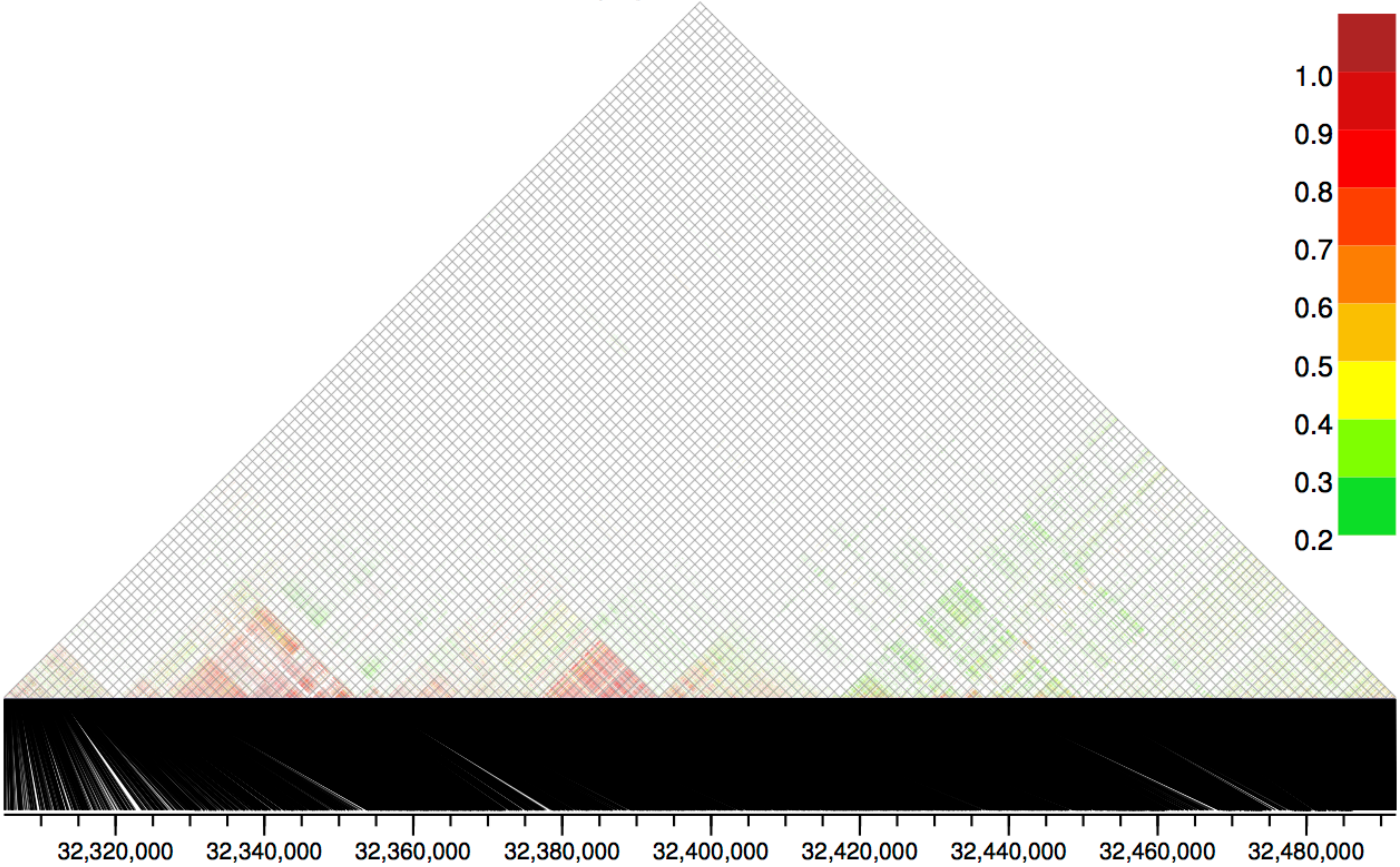
Problem statement

- Heatmap diagrams allow to quickly detect patterns that can be computationally verified and validated as statistically significant etc.
- In a glance, we can see many thousands of data items.
- However, when we visualize a table with thousands of rows we lose visibility of individual data items and of sparse patterns
- We overcome these limitations with aggregation and magnification

Examples

- Our examples use r^2 scores for African population computed in Kellis lab for HaploReg project with GRCh37 genomic coordinates.
- The first example is a region of 185 kbp with large LD blocks, the heatmap includes 5743 SNPs and added a grid, one line per 50 SNPs.

Chr6:32305000-32492000 5799 SNPs, 5743 displayed



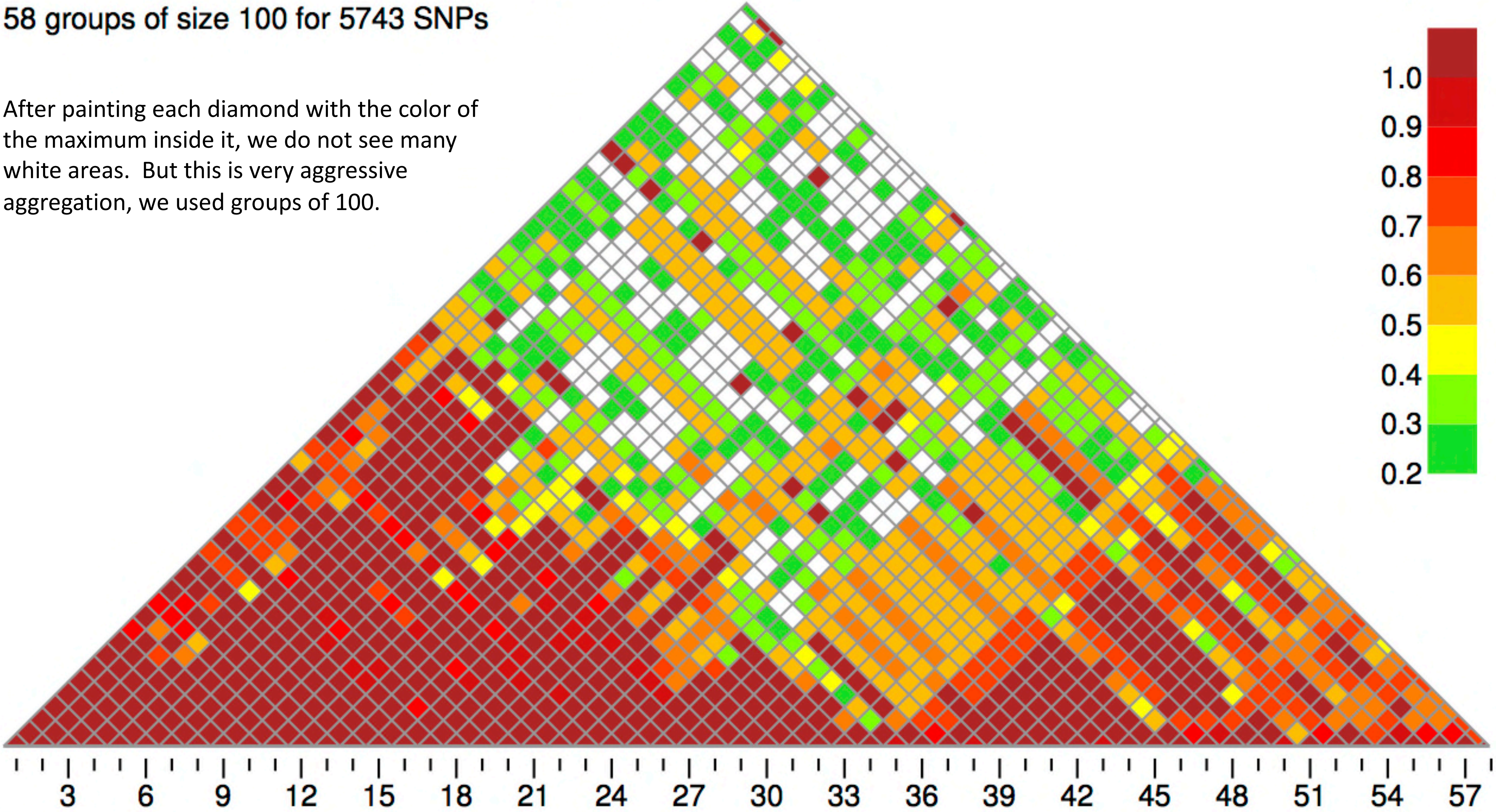
Identification of the problem

- The diagram is mostly white, and that could be expected because it visualizes ca. 15 million of scores of which less than 1 million exceed our reporting minimum of 0.2.
- But the visually “pure white” area contains many scores, and even some maximum scores, i.e. ones.
- How to find out if they exist? Aggregate!
- To aggregate, we will fill every grid diamond with the color of the maximum score in that diamond.

Chr6:32305000-32492000 5799 SNPs, 554443 scores

58 groups of size 100 for 5743 SNPs

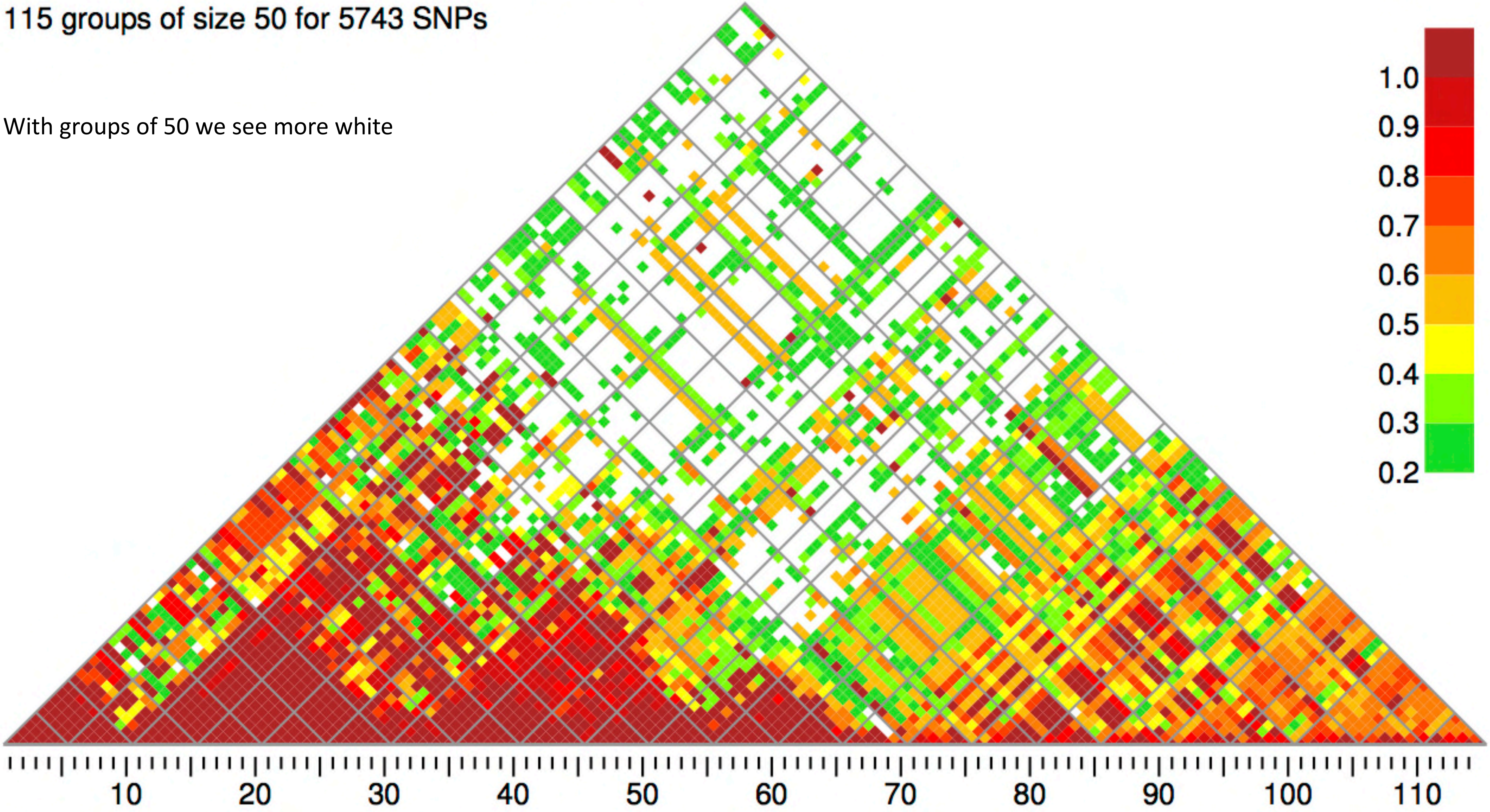
After painting each diamond with the color of the maximum inside it, we do not see many white areas. But this is very aggressive aggregation, we used groups of 100.

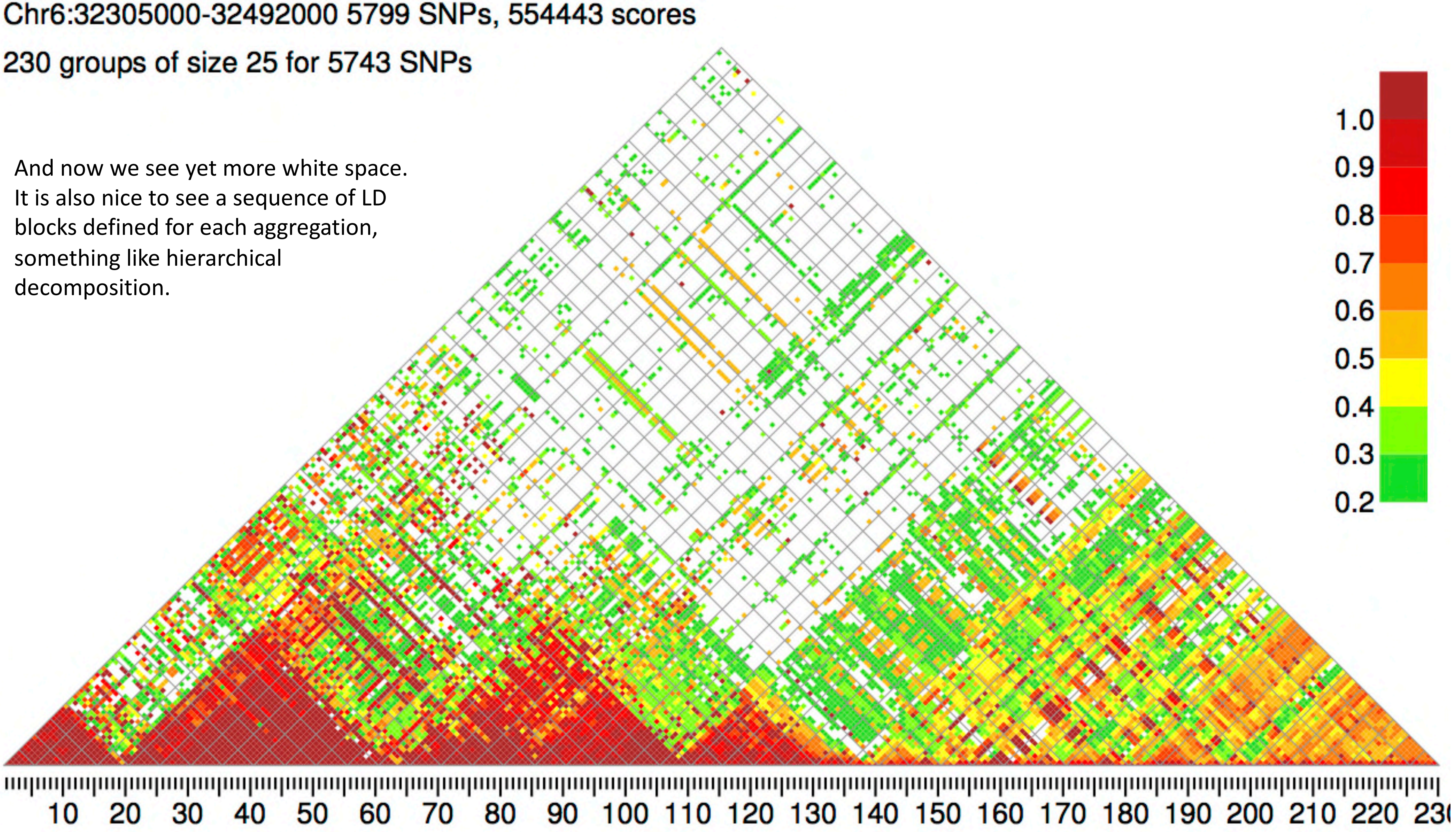


Chr6:32305000-32492000 5799 SNPs, 554443 scores

115 groups of size 50 for 5743 SNPs

With groups of 50 we see more white



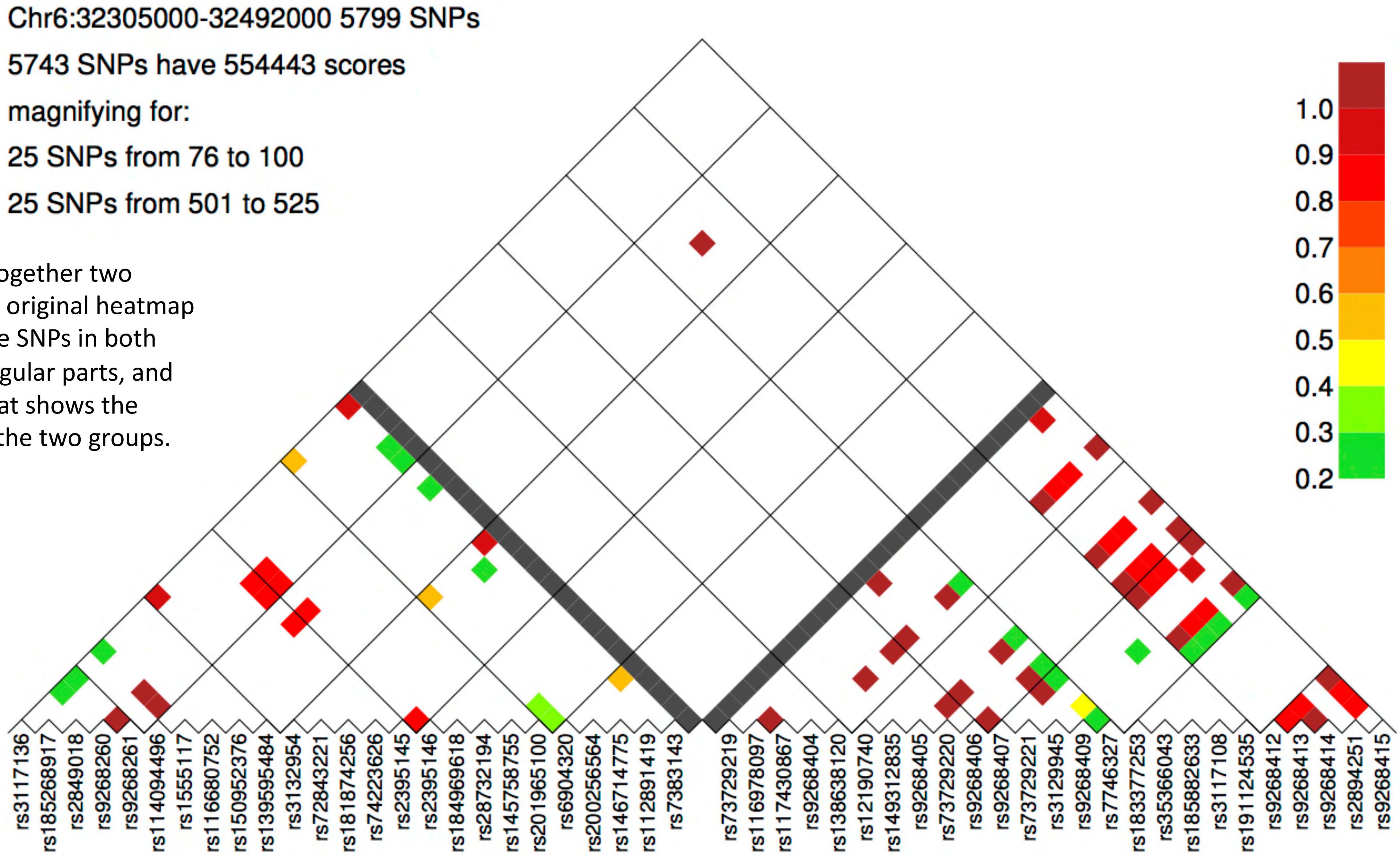


And now we see yet more white space. It is also nice to see a sequence of LD blocks defined for each aggregation, something like hierarchical decomposition.

Looking inside a little diamond -- magnify

- We can settle on the aggregation level that makes high scores visible but still provides a lot of detail
- Then if any of the colored diamonds is interesting, we can click on it to see the content of the respective fragment of the original heatmap.

We are putting together two fragments of the original heatmap that annotate the SNPs in both groups, the triangular parts, and the diamonds that shows the scores between the two groups.



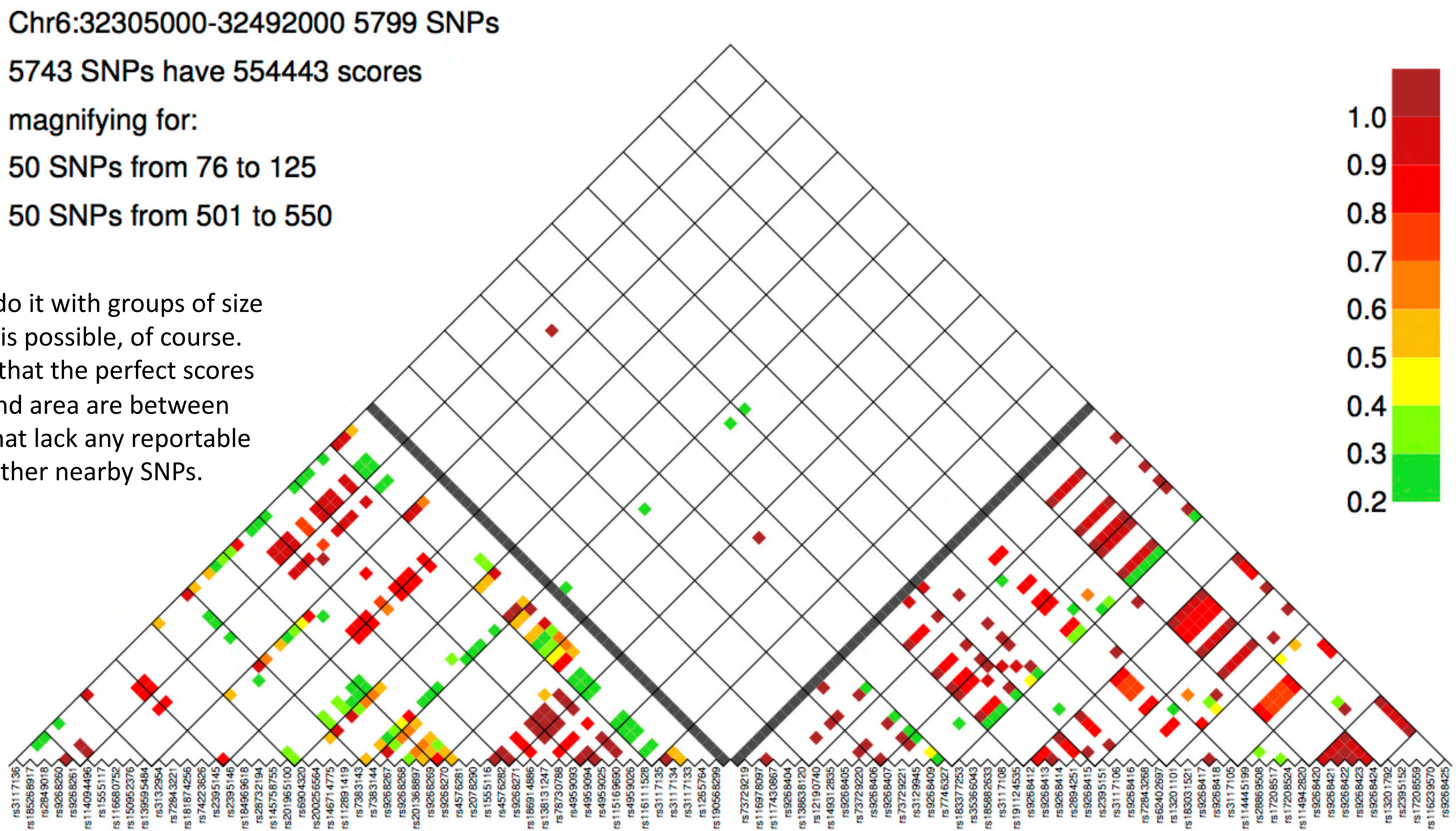
Chr6:32305000-32492000 5799 SNPs

5743 SNPs have 554443 scores

magnifying for:

25 SNPs from 76 to 100

25 SNPs from 501 to 525



We can also do it with groups of size 50 – any size is possible, of course. Here we see that the perfect scores in the diamond area are between three SNPs that lack any reportable scores with other nearby SNPs.

Conclusion

- Aggregation and magnification is easy for the user and provides a new valuable visual perspective on LD data