Detecting Inversions in Human Genome

Phillip Tao
Advisor: Eleazar Eskin

Abstract

The presence of inversion polymorphisms in the genome has been linked to increased likelihood of diseases such as Williams-Beuren Syndrome and colorectal cancer. However, because there is no widely established method for detecting inversions, we only know of a handful of token inversions in the human genome. This paper seeks to present a possible method for the detection of inversions using phased SNP (Single Nucleotide Polymorphism) data from the HapMap project.

1. Introduction

The genome is subject to a variety of mutations and structural variations, known as polymorphisms. These polymorphisms can be grouped into two general categories, single nucleotide polymorphisms, and multiple nucleotide polymorphisms. As the latter are much larger, a single instance of a deletion, duplication, or inversion can have a profound impact on the phenotype of the individual.

While much attention has been given by geneticists to SNPs, comparatively less research has been done on multiple base polymorphisms. This partially results from the fact that genotyping individuals is prohibitively expensive. However, with the success of the HapMap project, we now have a large database of SNP data which can be used to detect patterns consistent with the presence of inversions.

1.1 Key Concepts and Definitions

A **polymorphism** is a structural variation in the genome of an individual.

A **Single Nucleotide Polymorphism (SNP)** is a polymorphism at a single base pair.

Linkage Disequilibrium (LD) is the correlation between two SNPs.

An **inversion** is a section of the genome in which the order of the base pairs has been reversed.

A **breakpoint** is the location at either end of an inversion.

2. Background

In 2005, Lars Feuk et al. published a paper investigating the presence of human inversions by comparing the human genome alignment to that of a gorilla. 1,576 putative inversions were identified ranging from 23 bp to 64 Mb. However, this method does not take into account the possibilities of different human populations having differing inversions.

In 2006, Vikas Bansal et al. published a paper using the data from the HapMap project to detect inversions by looking for linkage disequilibrium (LD) patterns likely to result from an inversion. Using this method, they were able to identify 176 putative inversions ranging from 200 kb to 4 Mb.

This paper seeks to extend Bansal et. al's research by simplifying their methods.

3. Method

An inversion polymorphism leaves a unique pattern in the LD structure of a chromosome. Consider four base pairs, L_1, L_2, L_3, and L_4, with L_1 and L_2 on either side of the first breakpoint, and L_3 and L_4 on either side of the other breakpoint. We would expect L_1 and L_2 to be highly correlated, as they are close to each other, and likewise with L_3 and L_4. However, because of the inversion, L_1 and L_3 will actually be much more correlated than L_1 and L_2. This is because L_1 and L_3 were previously close to each other before the section was inverted, likewise for L_2 and L_4.

Therefore, by looking for such patterns in the phased HapMap data, we can detect possibly inverted sections. My algorithm does this in four steps.

3.1 Calculating the Correlation

The program first calculates the correlation between all SNPs. For each SNP, the alleles (180 in the CEU population and 134 in the JPTCHB population) are stored in binary with 0 representing the reference allele (A) and 1 representing the other allele (a). This data is then used to calculate P(a) for all SNPs. Then, the correlation between all SNPs is calculated using the formula for $|r|$.

$$|r| = \left| \frac{P_{1,2}(a,a) - P_1(a)P_2(b)}{\sqrt{P_1(a) * P_1(A) * P_2(a) * P_2(A)}} \right|$$

3.2 Calculating Difference

Next, for each SNP, we calculate the difference in the correlation between the other SNPs to itself. For each SNP, this is done in two parts.

First, you calculate the difference for all SNPs to the left of the given SNP. The difference is calculated by subtracting the correlation with the SNP to the furthest left from the correlation with the one closer to the given SNP.

Likewise, this is done for the SNPs on the right, subtracting the correlation with the furthest right from the one closer to the given SNP.

For both the left and right, the pair is only considered if the distance between the SNP closest to the given SNP and the given SNP is less than the distance between the two SNPs in the pair.

3.3 Identifying SNP Sets

Next, sets of 4 SNPs which fit the LD pattern described earlier are identified.

First, for each given SNP, look for all SNP pairs to the right which have a higher difference in correlation than a given threshold. These SNP pairs are stored in a table of right SNP pairs.

Then, for the same SNP, look for all SNP pairs to the left which have a significant difference in correlation. For each pair found, search through the table of right SNP pairs for a match.

If a match is found, the four SNPs are stored.

3.4 Grouping SNP Sets

Next, the sets of 4 SNPs are grouped based on position. Each group is saved as follows: L_{1L}, L_{1R}, L_{2L}, L_{2R}, L_{3L}, L_{3R}, L_{4L}, and L_{4R}, with L_{xL} being the left bound of L_x, and L_{xR} being the right bound. Each group also has a count of how many sets it is comprised of.

For each set of 4 SNPs, if there already exists a group such that both breakpoints match, that group's count is incremented, and the range of each SNP is expanded if needed.

A left breakpoint match is defined as L_1 being before L_{2L}, or L_2 being after L_{1R}. Likewise for the right breakpoint.
If needed, the bounds for L_x are expanded.

If there are no groups that match, a new group is created, with the left and right bound of each SNP being set to the SNP itself.

4. Results

I used the program to detect inversions in the first 8 ENCODE regions. The results can be found in the appendix. On average, about 4 putative inversions were found per ENCODE region per population. A total of 51 putative inversions were identified, with 12 being shared between the CEU and JPTCHB populations. The putative

inversions ranged between 7229 bp to 377 kb. The average length of inversions in the CEU population was 71.65 kb, whereas the average length of the inversions in the JPTCHB was 67.29 kb. The putative inversions in the JPTCHB population were concentrated around at around 25-75 kb. In both populations, the number of inversions inversely correlated with the length of the inversion. However, in both populations, after an initial cutoff of about 150 kb, there is a sudden spike at about 250 kb. A graph of the spread of lengths of putative inversions can be found in the appendix.

5. Conclusion

My method was moderately successful at identifying inversions. However, the third step, finding the difference in correlation, ran in $O(n^3)$ time, therefore the program is very slow. Another problem is the grouping algorithm used, it did not effectively group all the sets in the same position, as I still ended up with dozens of redundant groups for each group, which I had to pick through by hand. It also gave equal weight to all sets, when ideally it should weight sets in which the two inner SNPs are very far from each other and close to the two outer SNPs more than sets in which the two inner SNPs are very close. Also, as the ENCODE regions did not happen to overlap with any of the known human inversions, I was not able to verify any of my putative inversions.

Bibliography

[1] Feuk L, MacDonald JR, Tang T, Carson AR, Li M, Rao G, Khaja R, Scherer SW. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS Genet. 2005;1:e56. doi: 10.1371/journal.pgen.0010056.

[2] V. Bansal, A. Bashir, and V. Bafna. Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome Res., February 1, 2007; 17(2): 219 - 230.

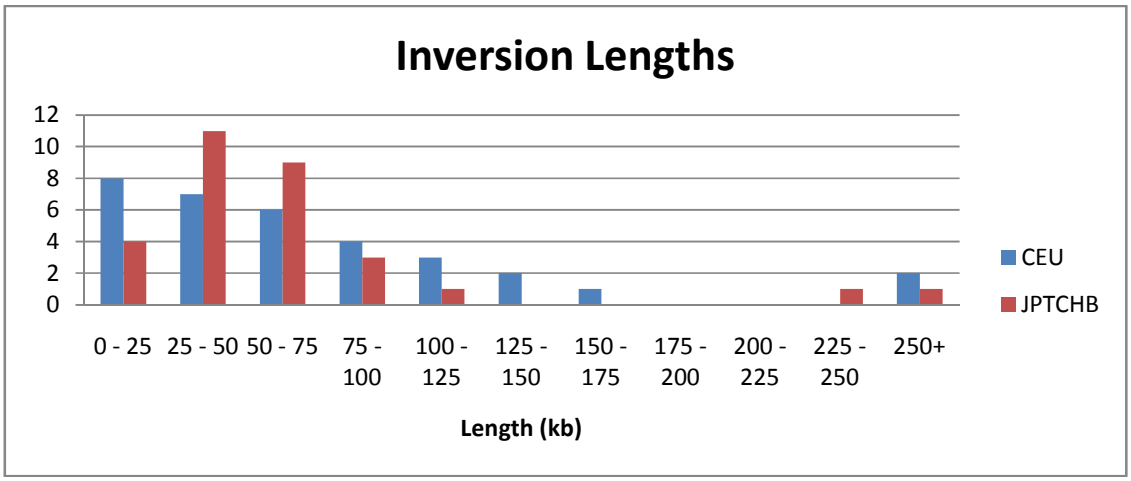
Acknowledgements

I would like to thank Dr. Eleazar Eskin for his direction and advice.
I would also like to thank Dr. Amit Sahai for the opportunity to do this research.

Appendix

Encode Region	CEU	CEU Length	JPTCHB	JPTCHB Length
1 (Chromosome 7)	**26935778 - 27073984**	**138206**	**26936185 - 27048950**	**112765**
	27112663 - 27125930	13267	27128499 - 27165385	36886
	27243243 - 27267966	24723		
	27292455 - 27309252	**16797**	**27292455 - 27309252**	**16797**
			27327781 - 27352782	25001
			27406545 - 27420237	13692
2 (Chromosome 7)	89703442 - 89850022	146580	89754595 - 89816526	61931
	89890370 - 90001689	111319	90071367 - 90448504	377137
	90037945 - 90141147	103202		
	90091549 - 90344508	252959		
	90155325 - 90235394	**80069**	**90153319 - 90251322**	**98003**
			90223524 - 90464701	241177
			90253143 - 90295959	42816
	90393602 - 90452814	**59212**	**90384689 - 90452814**	**68125**
3 (Chromosome 7)			126158247 - 126187950	29703
			126370050 - 126397326	27276
	126435292 - 126488603	**53311**	**126435292 - 126488603**	**53311**
	126552172 - 126630292	78120	126510624 - 126567343	56719
	126653273 - 126706661	53388		
	126731062 - 126810226	**79164**	**126749724 - 126821184**	**71460**
	126810226 - 126827248	17022		
4 (Chromosome 2)	51539277 - 51546506	7229		
	51556480 - 51598473	**41993**	**51540436 - 51588998**	**48562**
	51670492 - 51760902	90410	51644943 - 51701356	56413
			51788252 - 51873864	85612
5 (Chromosome 4)	118513984 - 118547958	33974		
	118649745 - 118694640	44895	118626418 - 118669353	42935
	118773291 - 118840239	66948		
	118865828 - 118940754	**74926**	**118877276 - 118944373**	**67097**
6 (Chromosome 12)	38781899 - 38829812	47913	38794431 - 38878607	84176
	38805440 - 39094939	289499		
	38912645 - 39090360	177715		
	39090653 - 39107194	**16541**	**39082553 - 39107443**	**24890**
7 (Chromosome 2)	234181821 - 234229144	47323		
	234286564 - 234319061	**32497**	**234290036 - 234321256**	**31220**
	234423903 - 234435866	11963	234428239 - 234453000	24761
	234468588 - 234519286	50698	234435292 - 234491387	56095
8 (Chromosome 18)	**23818749 - 23872998**	**54249**	**23825464 - 23876096**	**50632**
			23847929 - 23893571	45642
	23878593 - 23912905	**34312**	**23878593 - 23912905**	**34312**
	24049772 - 24063804	14032	24063033 - 24096604	33571

* bold entries are common between both populations


Inversion Lengths
12
10
8
6
4
2
0
0 - 25
25 - 50
50 - 75
75 - 100
100 - 125
125 - 150
150 - 175
175 - 200
200 - 225
225 - 250
250+
Length (kb)
CEU
JPTCHB