Preliminary data analytics on genomic breaks

Dario Garcia-Gasulla, in collaboration with David Torrents, Luisa Delgado, Juan Blanco Heredia, Armand Viltalta and Ferran Parés.

February 22, 2018

Data overview

Data is obtained from 2,784 patients, with the corresponding breaks. All patients are considered as independent instances, and all breaks as independent variables. Each break has a source or target, which simply identifies an arbitrary order of break. There are 5 types of breaks as provided by the .vcf.tsv files: h2hINV, DEL, t2tINV, DUP, TRA.

Chromosome break distribution

Distribution of breaks

First we plot the number of breaks per chromosome, regardless of type.

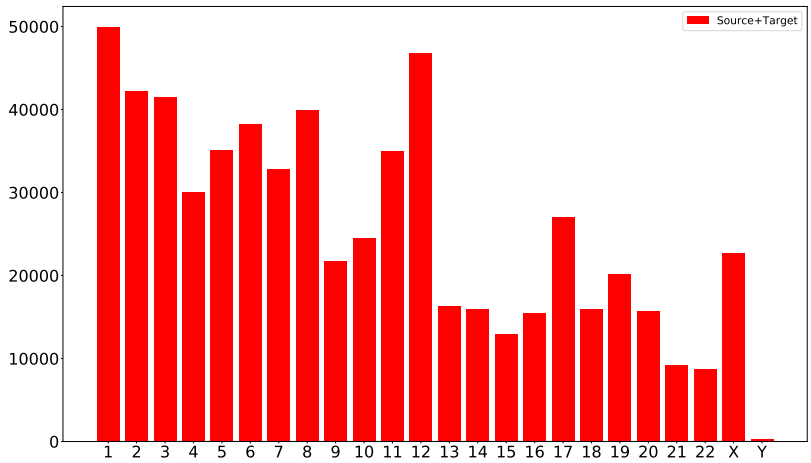

Figure 1: Distribution of breaks per chromosome. All brake types are aggregated. Values not normalized

We normalize by chromosome length, using the following lengths (from https://en.wikipedia.org/wiki/Human_genome):

- chromosome_size['1'] = 248956422.0
- chromosome_size['2'] = 242193529.0
- chromosome_size['3'] = 198295559.0
- chromosome_size['4'] = 190214555.0
- chromosome_size['5'] = 181538259.0
- chromosome_size['6'] = 170805979.0
- chromosome_size['7'] = 159345973.0
- chromosome_size['8'] = 145138636.0
- chromosome_size['9'] = 138394717.0
- chromosome_size['10'] = 133797422.0
- chromosome_size['11'] = 135086622.0
- chromosome_size['12'] = 133275309.0
- chromosome_size['13'] = 114364328.0
- chromosome_size['14'] = 107043718.0
- chromosome_size['15'] = 101991189.0
- chromosome_size['16'] = 90338345.0
- chromosome_size['17'] = 83257441.0
- chromosome_size['18'] = 80373285.0
- chromosome_size['19'] = 58617616.0
- chromosome_size['20'] = 64444167.0
- chromosome_size['21'] = 46709983.0
- chromosome_size['22'] = 50818468.0
- chromosome_size['X'] = 156040895.0
- chromosome_size['Y'] = 57227415.0

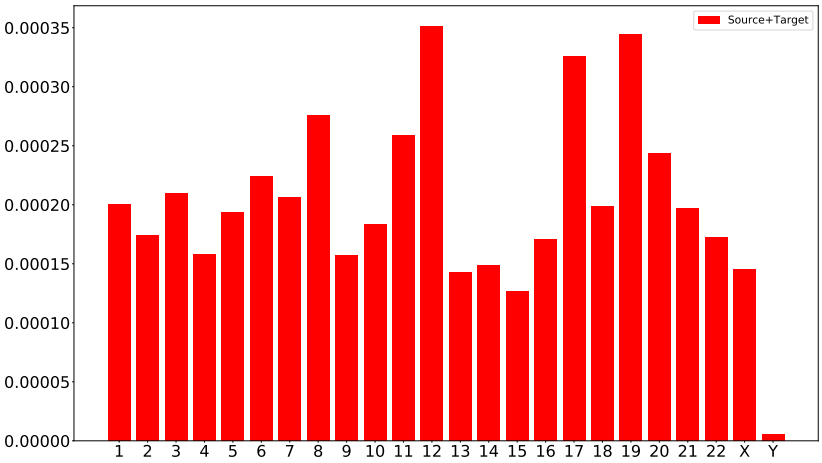

Figure 2: Distribution of breaks per chromosome. All brake types are aggregated. Values normalized by chromosome length

Distribution of breaks by type, unnormalized

Next we plot the distribution of breaks by break type (h2hINV, DEL, t2tINV, DUP, TRA), unnormalized

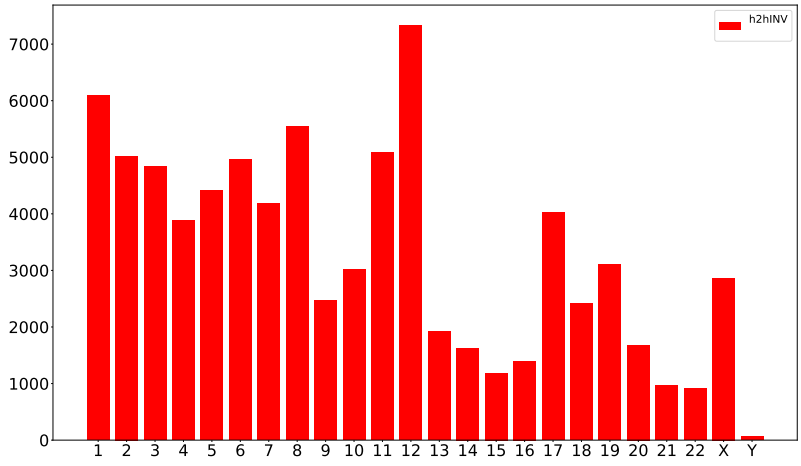

Figure 3: Distribution of h2hINV breaks per chromosome. Values unnormalized

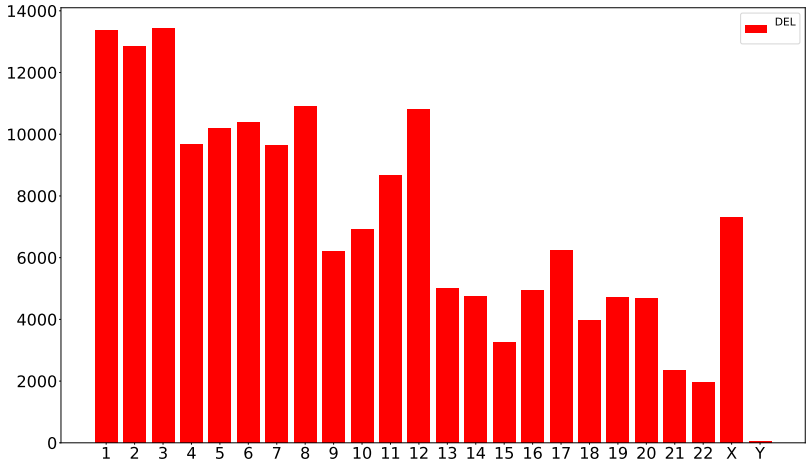

Figure 4: Distribution of DEL breaks per chromosome. Values unnormalized

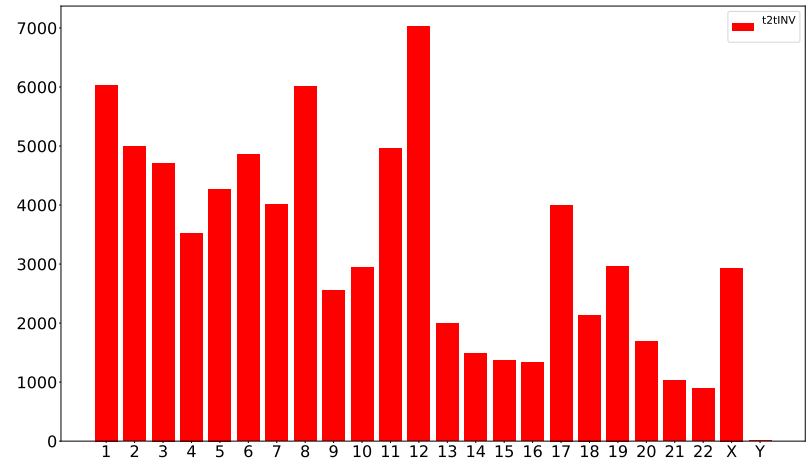

Figure 5: Distribution of t2tINV breaks per chromosome. Values unnormalized

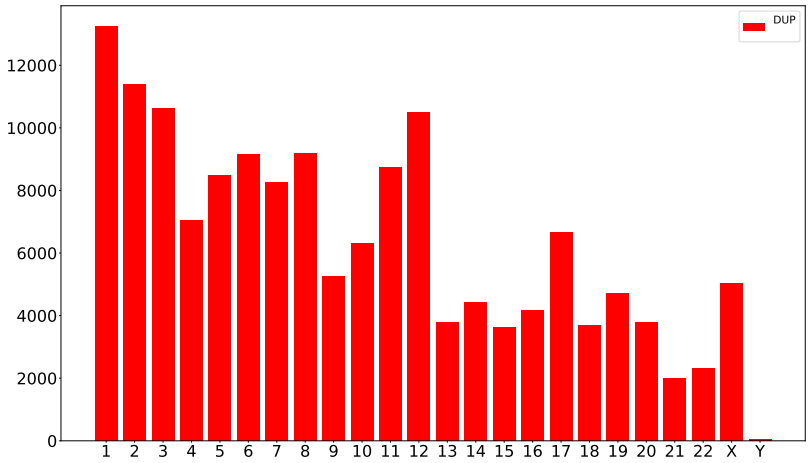

Figure 6: Distribution of DUP breaks per chromosome. Values unnormalized

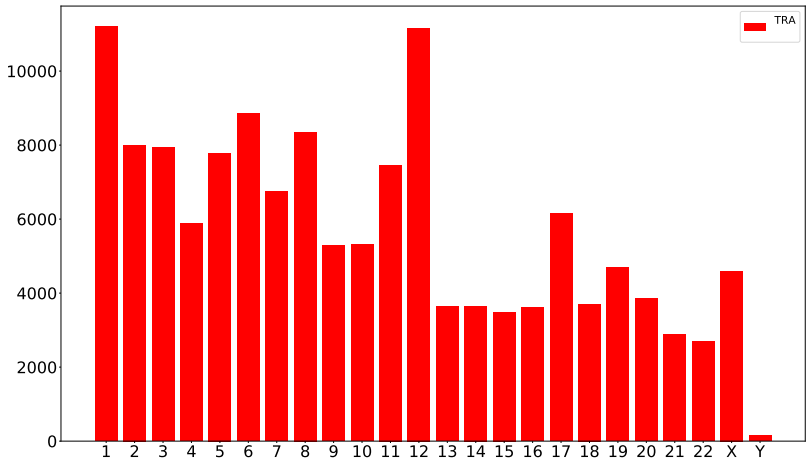

Figure 7: Distribution of TRA breaks per chromosome. Values unnormalized

Distribution of breaks by type, normalized

Next we plot the distribution of breaks by break type (h2hINV, DEL, t2tINV, DUP, TRA), normalized by chromosome length

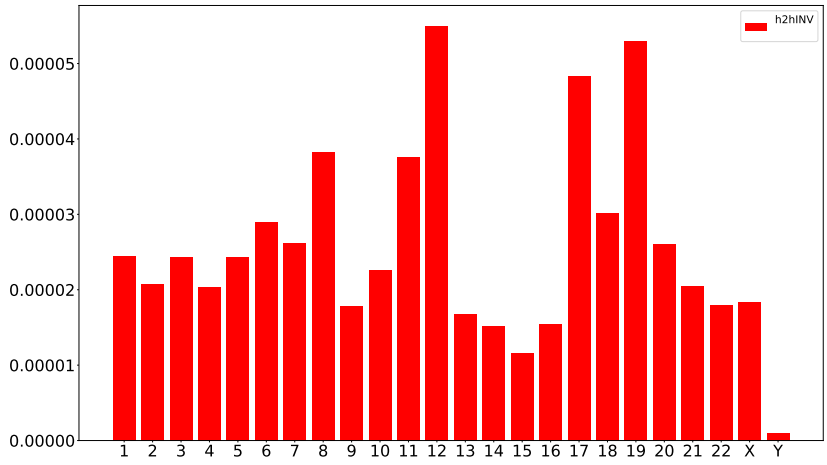

Figure 8: Distribution of h2hINV breaks per chromosome. Values normalized by chromosome length

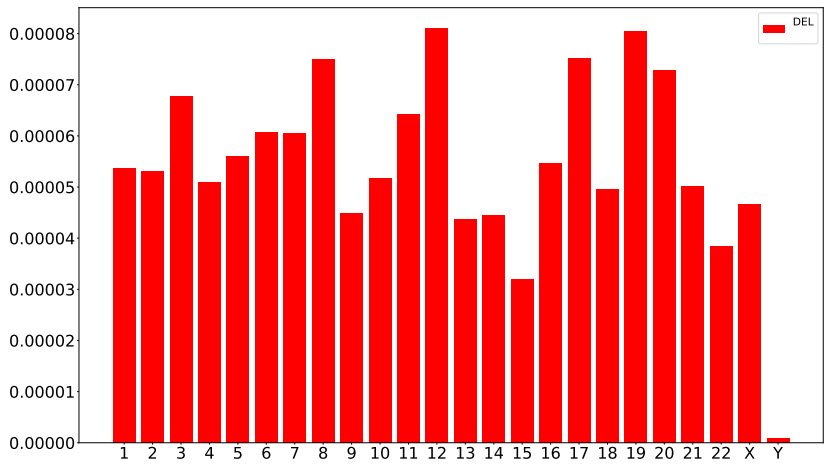

Figure 9: Distribution of DEL breaks per chromosome. Values normalized by chromosome length

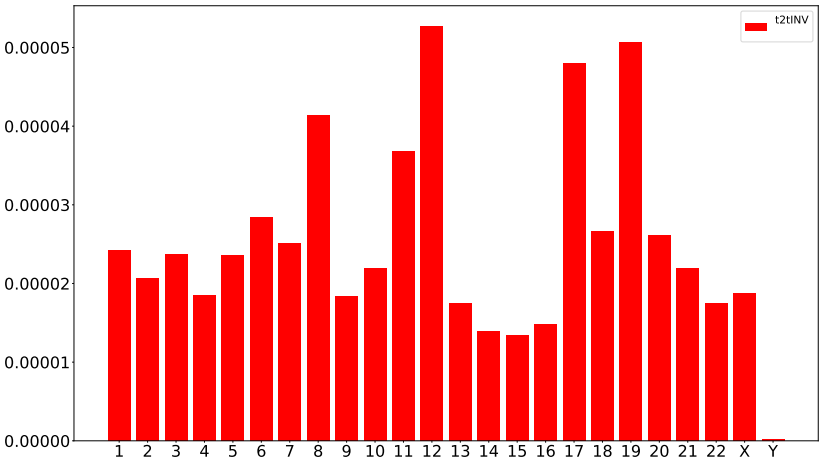

Figure 10: Distribution of t2tINV breaks per chromosome. Values normalized by chromosome length

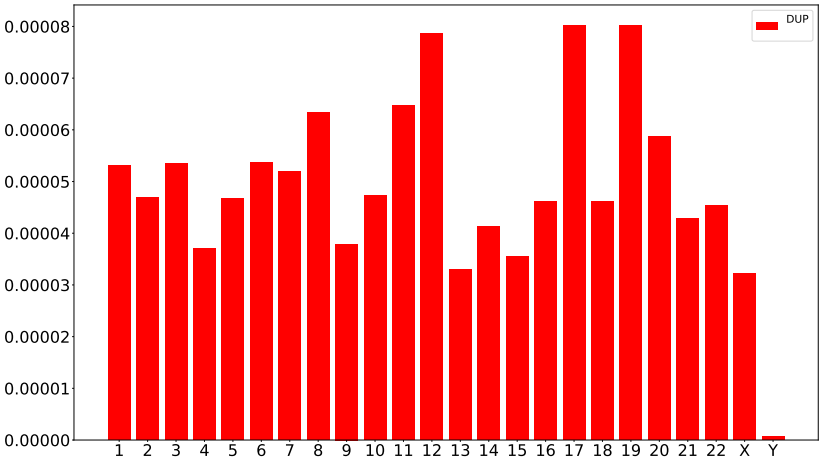

Figure 11: Distribution of DUP breaks per chromosome. Values normalized by chromosome length

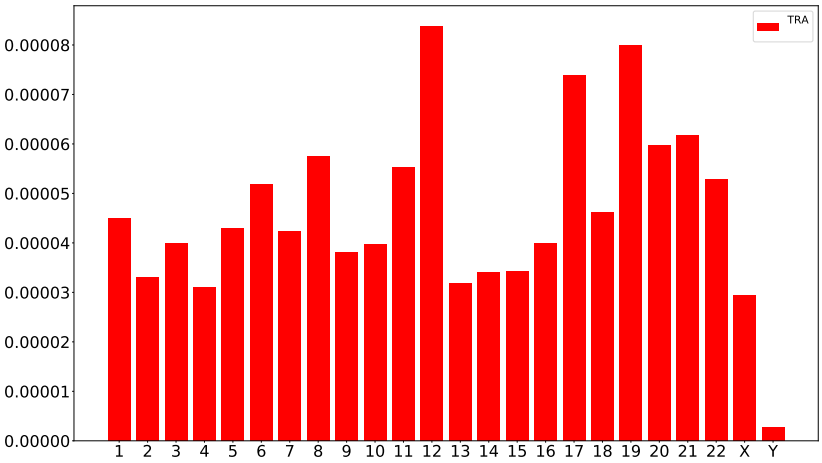

Figure 12: Distribution of TRA breaks per chromosome. Values normalized by chromosome length

Heat map of break interaction

To find the most common break pairs, we plot a matrix of frequencies, like a heatmap.

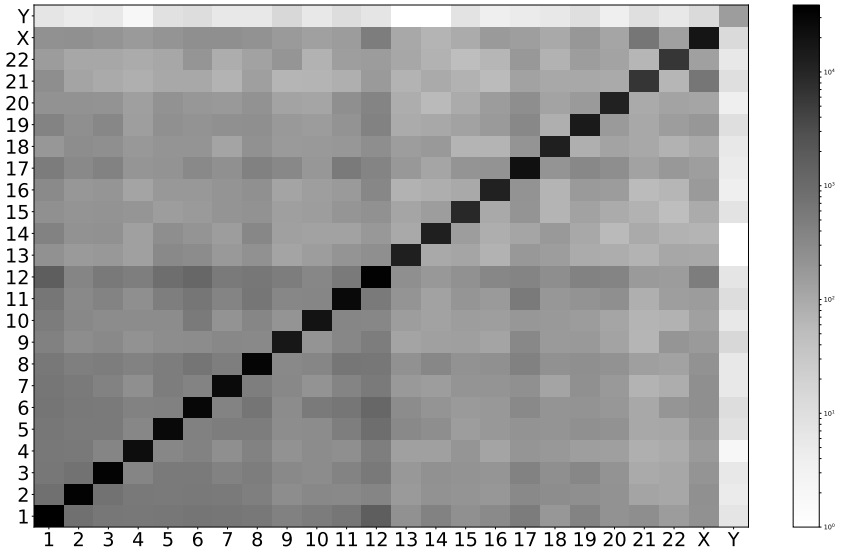

Figure 13: Matrix of break frequencies among chromosomes. Values unnormalized. Logarithmic scale.

To clarify, we remove the diagonal, which is dominating the coloring.

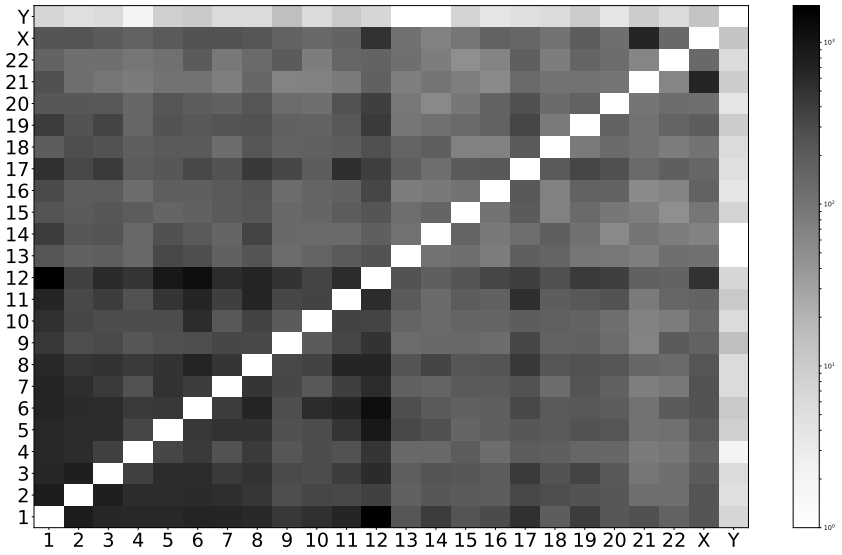

Figure 14: Matrix of break frequencies among chromosomes. Values unnormalized. Logarithmic scale. Diagonal removed.

Heat map of break interaction, normalized

We normalize the frequencies by the length of both chromosomes involved

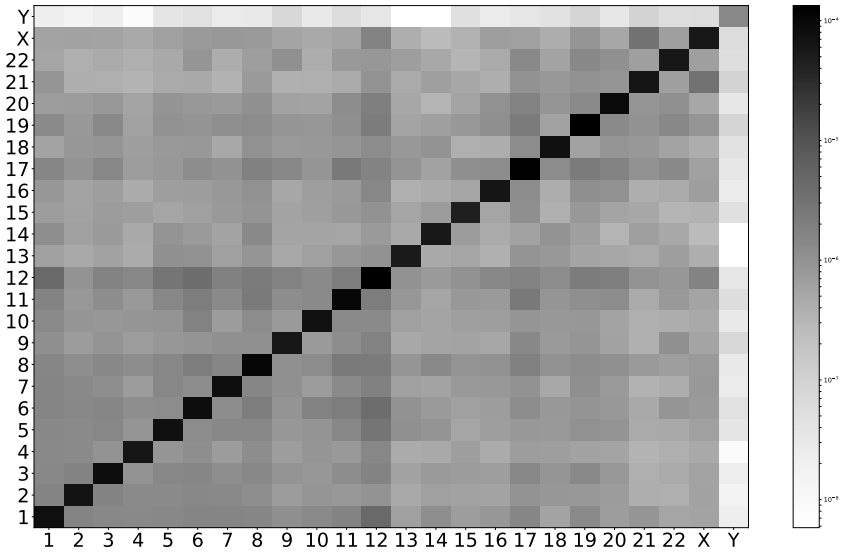

Figure 15: Matrix of break frequencies among chromosomes. Values normalized by length of the two chromsomes. Logarithmic scale.

To clarify, we remove the diagonal, which is dominating the coloring.

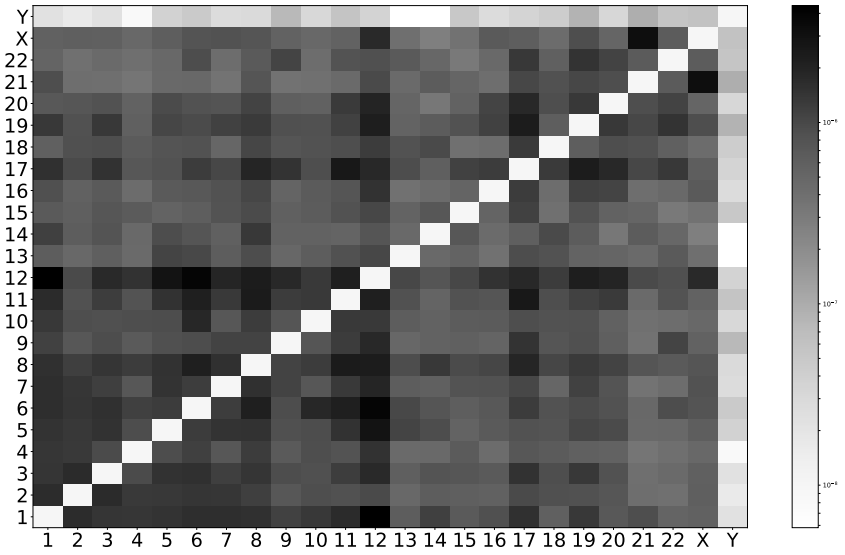

Figure 16: Matrix of break frequencies among chromosomes. Values normalized by length of the two chromsomes. Logarithmic scale. Diagonal removed.

For reference, we list the top frequencies, avoiding the diagonal:

1. 11 12
2. 20 12
3. 12 7
4. 17 8
5. 10 6
6. 12 9
7. 20 17
8. 12 17
9. 12 23
10. 3 12
11. 1 11
12. 2 3
13. 1 2
14. 6 1

15. 1 7
16. 8 1
17. 7 8
18. 1 17
19. 3 6
20. 8 5
21. 11 5
22. 3 17
23. 16 12
24. 5 3
25. 12 4
26. 7 5
27. 9 17
28. 1 5
29. 22 19