Supplement to: The Genomic Sequence of the Chinese Hamster Ovary (CHO)-K1 cell line

Table of Contents

Nature Biotechnology: doi:10.1038/nbt.1932

Supplementary Text:

Filtering of raw reads prior to assembly:

We filtered the Illumina reads based on several criteria such as:

- Filter reads where more than 10% of the bases were degenerate (N).
- Quality filtering of the reads with a Q40 cutoff.
- Filter for contamination caused by DNA adaptor dimers.
- Filter reads which were PCR duplicates (Two totally identical reads).

Comparative analysis

Many of the predicted 24,383 genes have homologs in human (19,711), mouse (20,612) and rat (21,229). There were 3,433 mouse genes (consisting largely of transmembrane proteins and odorant receptors) without a homolog in the CHO genome. Orthologous clusters among the CHO, human, mouse and rat genomes were computed using OrthoMCL[1]. We found 15,851 CHO genes grouped into 13,571 orthologous clusters, out of which 407 genes were CHO-K1 specific (i.e., not shared with either human, mouse or rat). 13,374 CHO genes have orthologs in human, mouse and rat. CHO-K1 shares 862 genes with mouse and rat, while only sharing 596 genes with human and mouse and 177 with human and rat. This result suggests that, at the genome-level, CHO-K1 more closely resembles the rodents than human (Supplementary Figure 4). Lastly, 2,428 gene clusters contained human, mouse and rat genes but no CHO-K1 genes. These clusters are enriched for genes involved in defense and immune response (Supplementary Table 12).

Immunogenic responses due to glycan structures:

Variations in glycan structure between individuals can also result from the common ABO blood group system. These glycans include the H blood group structure (O blood type), and the A and B structures that result from different alleles of the ABO gene. In addition to ABO, the H and Secretor loci encode additional glycosyltransferases that are responsible for the various glycan structures. In the synthesis of the ABO antigens, a precursor glycan is modified by α1–2 fucosyltransferases, encoded by the H or Secretor genes (FUT1 and FUT2). These enzymes form a disaccharide unit of Fucα1–2Galβ1-, which is the H structure. This structure can be further modified by the glycosyltransferase encoded by the ABO gene. It has been shown that the H and Secretor gene products have no detectable activity in CHO cells [2]. Moreover, while we find homologs for these genes in the CHO-K1 genome, there are no detectable levels of expression for either of these genes as well as the ABO gene. This further limits the production of proteins with potential immunogenic responses in humans.

Identification of non-coding RNAs

Non-coding RNAs: Non-coding transcripts do not create protein products; while many remain functionally uncharacterized, some of which may have novel functions. Sequenced transcripts from a pooled CHO sample were compared to the fRNAdb database of non-coding RNAs[3] using BLAST. Although the entire fRNAdb consists of sequences from multiple species (mainly human, mouse, rat and zebra fish), most of the top alignment hits for long ncRNAs were those of mouse, corroborating pervious results demonstrating the similarities between CHO and mouse genomes (see Supplementary Table 7 for the ncRNAs).

A breakdown of high scoring sequences into different ncRNA classes was examined; particularly pre-miRNAs which are the precursors of single-stranded (ss) miRNAs whose expression in animals and plants are known to affect the translation or stability of target mRNAs. Most of these pre-miRNAs that CHO transcripts aligned originated mainly from mouse and rat (see Supplementary Table 8).

snoRNAs are a class of small RNA molecules that primarily guide chemical modifications of other RNAs. The majority of snoRNAs to which CHO sequences align belong to a particular class of Small Cajal body-specific RNAs (scaRNAs) from mouse and human. scaRNAs localize to Cajal bodies, which are sub-organelles typically found in the nucleus of proliferative and metabolically active cells like embryonic or tumor cells. See Supplementary Table 9 for identified snoRNAs.

Assignment of scaffolds to chromosomes

Reads from each of the chromosome preparation were aligned to the draft genome reference using the BWA[4] package. Subsequently, SAMtools[5] were used to filter for high quality mapping of reads with score of greater than 30. Read alignment files were then parsed using BEDtools[6] to get the number of reads that are associated with each of the scaffolds. The number of reads associated with each scaffold from a particular chromosome preparation were arranged in the following data format. In each cell the number of reads associated with the scaffold and preparation is normalised by the total number of reads across each column so that a valid comparison can be made across each preparation. Normalisation is performed across the column by dividing each entry by the total number of reads of the column.

Names of Scaffold	**Chromosome Preparation 1**	**…**	**Chromosome Preparation n**
scaffold 1	Number of reads associated with the scaffold and preparation	…	…
…	…	…	…
Scaffold n	…	…	…

To determine the assignment of the scaffold to a particular chromosome preparation, two measures were used. First, the maximum number of normalized reads associated with each scaffold is divided by the sum of the normalized reads across the row *i*.

$$x_i = \frac{max_i(row\ i)}{\sum_1^{Length\ of\ row} row\ i}$$

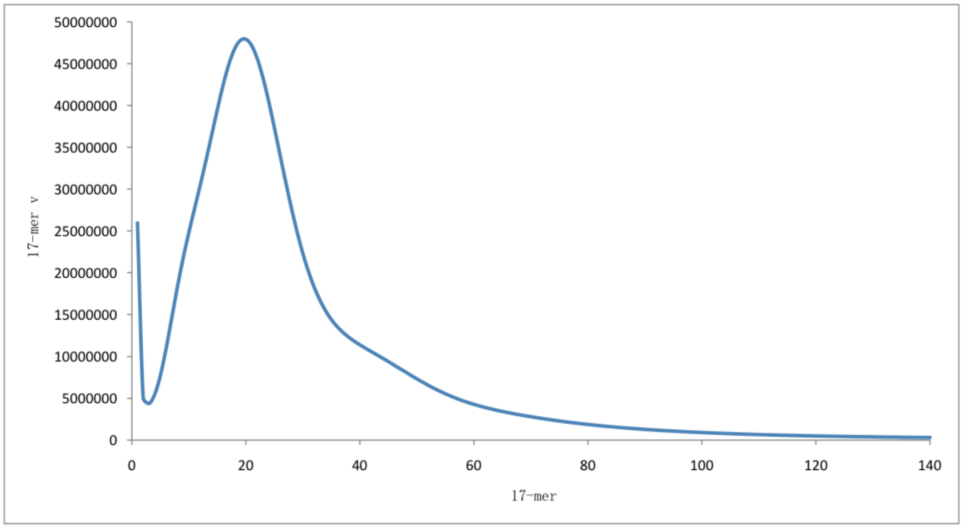
Second, the second largest number of the normalized reads associated with each scaffold is divided by the sum of the normalized reads across the row *i.*

$$y_{i=} = \frac{2nd\ max_i(row\ i)}{\sum_1^{Length\ of\ row} row\ i}$$

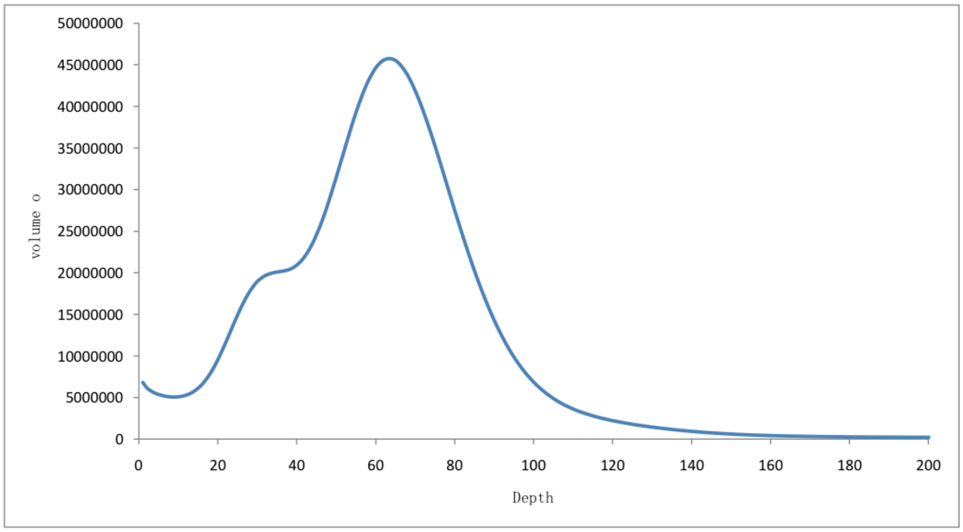
For each of the scaffolds, the associated x and y values are plotted. The values of x and y are mathematically bounded by the following inequalities: x>y (red line); x+y<1 (blue line) ; 0<x<1 ; 0<y<1 (Supplementary Figure 5). To assign the chromosome with high stringency, a quantitative measure is calculated (i.e., fold change of the value x over the value y). This measure can be interpreted as fold change of the maximum normalized score for each scaffold against the second maximum score. A high fold change value will indicate a strong confidence that the scaffold belongs almost uniquely to a particular chromosome preparation and will be assigned the respective chromosome preparation label. See Supplementary Table 19 for the values corresponding to the N50 scaffold in each of the chromosomal preparations.

The heatmap of the correlation matrix generated between all the N50 scaffolds suggested 21 clusters of highly correlated scaffolds. To gain further insight into the number of read associated with each scaffold across different sample preparations, K-means clustering was performed on the same matrix to label and separate the scaffolds into different groups. Individual heatmaps for each group were generated (Supplementary Figure 6). The X-axis is the samples label from S1 to S22 and the Y axis are the names of the scaffolds belonging to the group as determined from the clustering step. We obtained 21 group labels for the scaffolds, where a clear red vertical line indicates that a large number of sequenced reads mapping to those scaffolds originates from a particular single sample preparation (Supplementary Table 20).

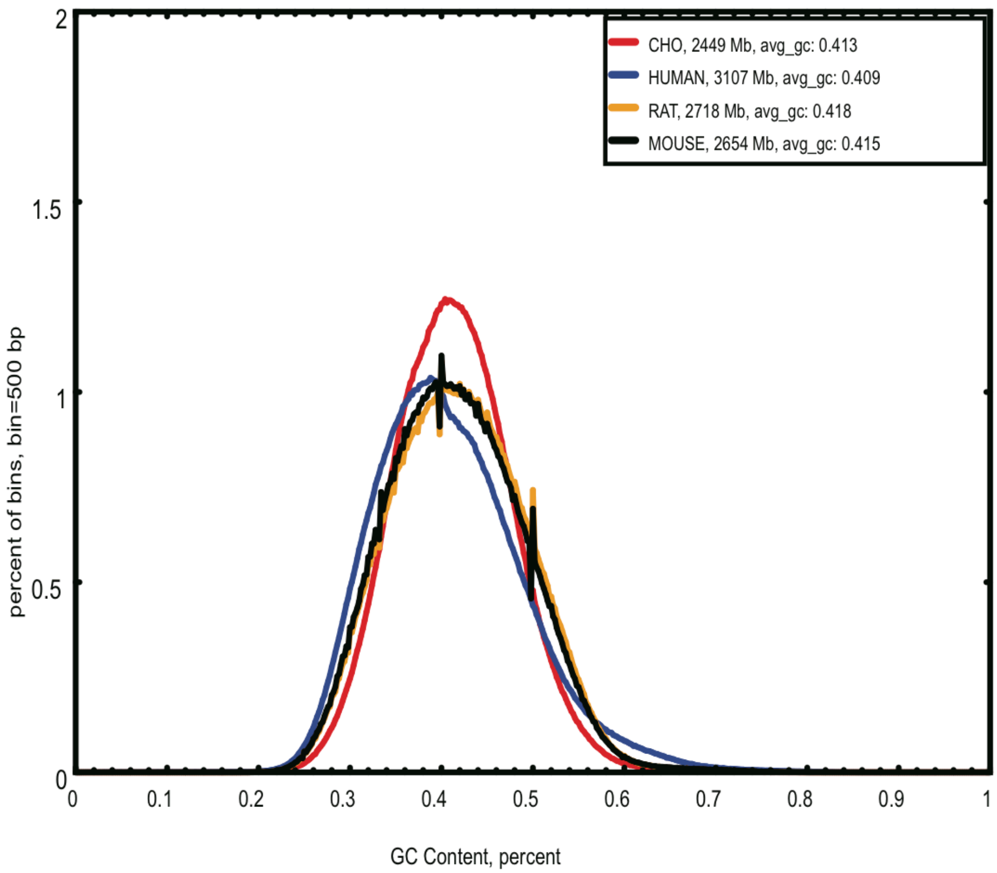
Supplementary Figures



Supplementary Figure 1:
Illumina 17-mer volume histograms of CHO-K1. The volume of K-mers (Y axis) is plotted against their frequency. The truncated peak at low frequency represents K-mers containing random sequencing errors. The volume peak is at 20 and the total K-mer number is 52,036,421,291. The estimated genome size (total K-mer number/volume peak) thus is 2.6 Gb.

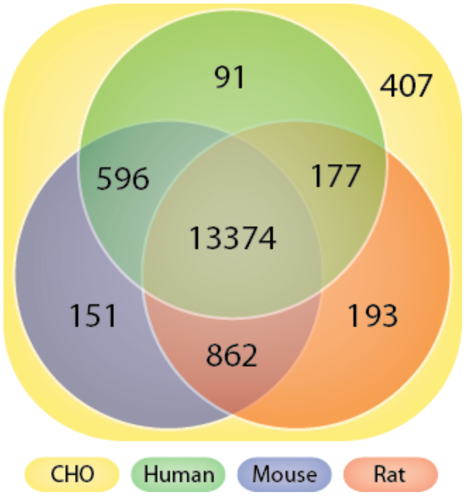


Supplementary Figure 2:
Sequencing depth distribution. Average sequencing depth is 63.29 and peak depth is 63.



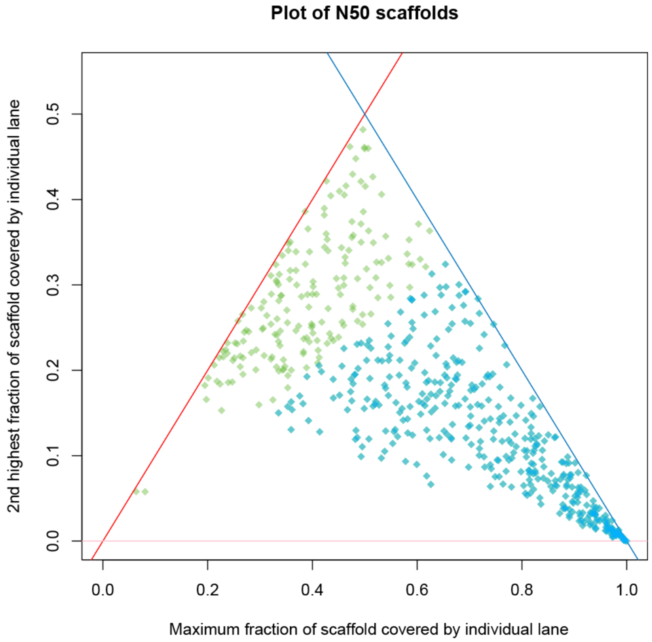
Supplementary Figure 3:
GC content distribution among CHO, mouse, rat and human. The X-axis is the GC percentage in the 500bp window and the Y-axis is the percentage of window with certain GC percentage (number of windows with certain GC percentage/total windows).

Nature Biotechnology: doi:10.1038/nbt.1932



Supplementary Figure 4

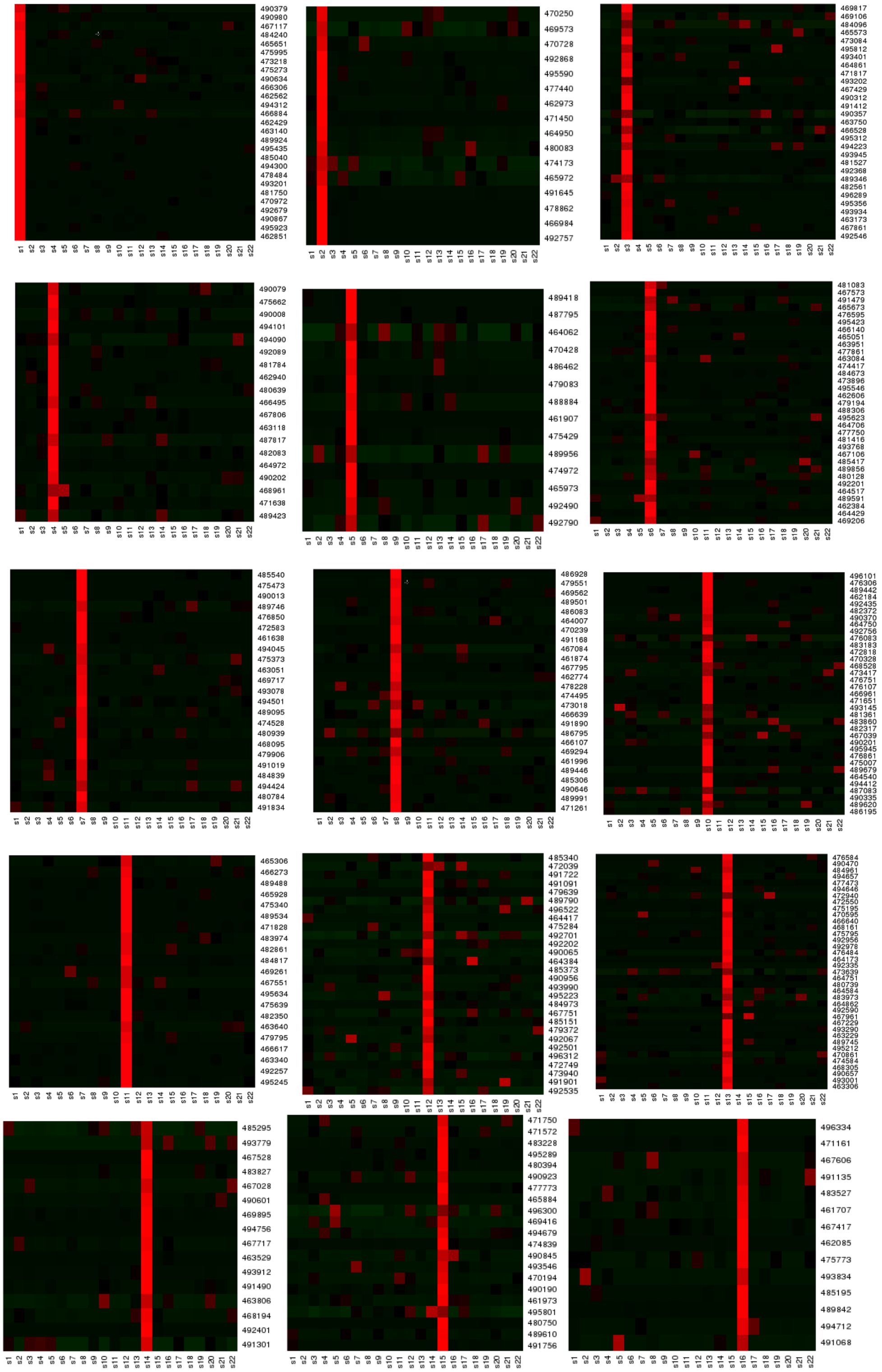
Orthologous clusters between the CHO, human, mouse and rat genomes show a high overlap of gene content between the organisms. A total of 15,851 CHO genes form orthologous clusters with human, mouse and rat genes. Only the number of CHO genes in each category is shown. The clusters without a CHO gene are not shown.

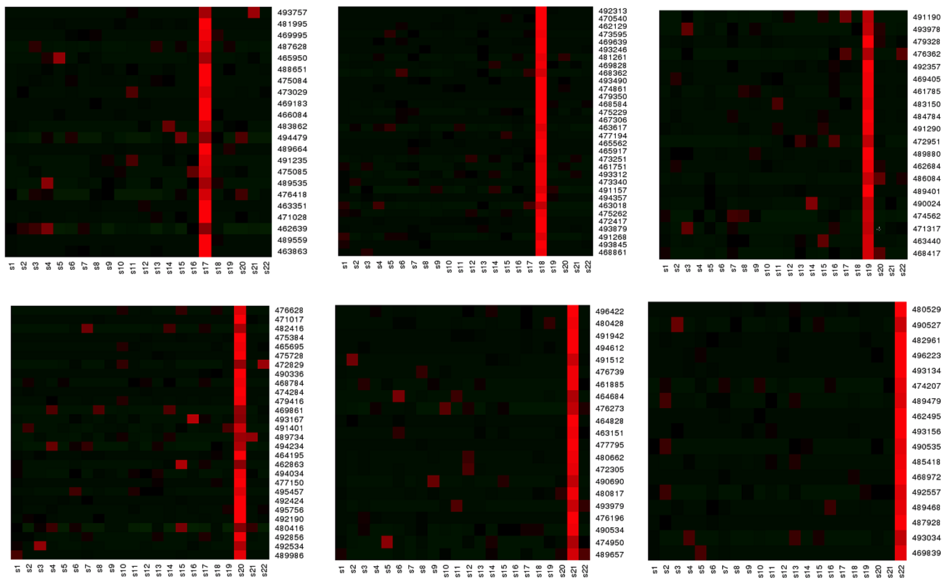


Supplementary Figure 5:

A scatter plot of the top two relative frequencies was used to identify scaffolds clearly corresponding to unique chromosome amplification. The red and blue lines represent the mathematical boundaries originating from the method of counting the frequency. Each point on the plot represents a scaffold from the set of N50 scaffolds, and each scaffold point is defined by two values, the maximum and the second largest fraction of the scaffold covered by an individual chromosome lane preparation. Points that are indicated in light blue represent a threshold that requires the maximum to be more than twice than that of the second largest value.

Nature Biotechnology: doi:10.1038/nbt.1932


490379 490980 467117 484240 465651 475995 473218 475273 490634 466306 462562 494312 466884 462429 463140 489924 495435 485040 494300 478484 493201 481750 470972 492679 490867 495923 462851
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
470250 469573 470728 492868 495590 477440 462973 471450 464950 480083 474173 465972 491645 478862 466984 492757
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
469817 469106 484096 465573 473084 495812 493401 464861 471817 493202 467429 490312 491412 490357 463750 466528 495312 494223 493945 481527 492368 489346 482561 496289 495356 493934 463173 467861 492546
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
490079 475662 490008 494101 494090 492089 481784 462940 480639 466495 467806 463118 487817 482083 464972 490202 468961 471638 489423
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
489418 487795 464062 470428 486462 479083 488884 461907 475429 489956 474972 465973 492490 492790
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
481083 467573 491479 465673 476595 495423 466140 465051 463951 477861 463084 474417 484673 473896 495546 462606 479194 488306 495623 464706 477750 481416 493768 467106 485417 489856 480128 492201 464517 489591 462384 464429 469206
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
485540 475473 490013 489746 476850 472583 461638 494045 475373 463051 469717 493078 494501 489095 474528 480939 468095 479906 491019 484839 494424 480784 491834
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
486928 479551 469562 489501 486083 464007 470239 491168 467084 461874 467795 462774 478228 474495 473018 466639 491890 486795 466107 469294 461996 489446 485306 490646 489991 471261
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
496101 476306 489442 462184 492435 482372 490370 464750 492756 476083 483183 472818 470328 468528 473417 476751 476107 466961 471651 493145 481361 483860 482317 467039 490201 495945 476861 475007 489679 464540 494412 487083 490335 489620 486195
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
465306 466273 489488 465928 475340 489534 471828 483974 482861 484817 469261 467551 495634 475639 482350 463640 479795 466617 463340 492257 495245
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
485340 472039 491722 491091 479639 489790 496522 464417 475284 492701 492202 490065 464384 485373 490956 493990 495223 484973 467751 485151 479372 492067 492501 496312 472749 473940 491901 492535
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
476584 490470 484961 494657 477473 494846 472940 472550 475195 470595 466640 468161 475795 492956 492978 476484 464173 492335 473639 464751 480739 464584 483973 464862 492590 467961 467229 493290 463229 489745 495212 470861 474584 468305 490657 493001 463306
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
485295 493779 467528 483827 467028 490601 469895 494756 467717 463529 493912 491490 463806 468194 492401 491301
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
471750 471572 483228 495289 480394 490923 477773 465884 496300 469416 494679 474839 490845 493546 470194 490190 461973 495801 480750 489610 491756
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22
496334 471161 467606 491135 483527 461707 467417 462085 475773 493834 485195 489842 494712 491068
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22



Supplementary Figure 6:

Collection of 17 groups of heatmaps from the K-means clustering analysis, which show the presence of 21 distinct chromosome groups (vertical lines).

References:

1. Chen, F., Mackey, A.J., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**, D363-368 (2006).
2. Prieto, P.A. et al. Expression of human H-type alpha1,2-fucosyltransferase encoding for blood group H(O) antigen in Chinese hamster ovary cells. Evidence for preferential fucosylation and truncation of polylactosamine sequences. *J Biol Chem* **272**, 2089-2097 (1997).
3. Kin, T. et al. fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res* **35**, D145-148 (2007).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
5. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
6. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).