Supplementary Data

i-ADHoRe 3.0 – Fast and Sensitive Detection of Genomic Homology in Extremely Large Data Sets.

Sebastian Proost\textsuperscript{1,2,4}, Jan Fostier\textsuperscript{3,4}, Dieter De Witte\textsuperscript{3}, Bart Dhoedt\textsuperscript{3}, Piet Demeester\textsuperscript{3}, Yves Van de Peer\textsuperscript{1,2,5} and Klaas Vandepoele\textsuperscript{1,2}

\textsuperscript{1} Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium;
\textsuperscript{2} Department of Plant Biotechnology and Genetics, Ghent University, B-9052 Ghent, Belgium;
\textsuperscript{3} Department of Information Technology - IBBT, Ghent University, B-9050 Ghent, Belgium

\textsuperscript{4} These authors contributed equally to this work.
\textsuperscript{5} Corresponding author.

E-mail yves.vandepeer@psb.vib-ugent.be
Supplementary Methods

Supplementary Method 1: Synteny Mode

Whereas collinearity is excellent to detect remnants of relatively recent duplications and homologous regions between closely related species, more ancient homologous regions may remain undetected (9). Synteny is a valid, albeit less stringent, alternative than collinearity to detect ancient homology between regions that experienced severe rearrangements, such as, for example, paralogous regions that originated from the whole-genome duplication (WGD) in the common ancestor of all vertebrates 350-450 Mya (20). i-ADHoRe 3.0 features an additional clustering algorithm to detect genomic homology based on shared gene content, coined the synteny mode.

The input and initial steps were identical to the collinearity search mode. Only after the pairwise GHMs had been built, a different clustering algorithm was used (Supplementary Figure 1). Clouds of dots contained by a bounding box were detected. Initially, the method started by considering all the dots of the GHM as potential cloud seeds. Subsequently the seeding algorithm searched a rectangular area, defined by the cloud_gap, for additional dots, and all dots in this window formed the seed cloud. Next, all seed clouds would grow by adding all dots present in a frame with a thickness equal to the cloud_gap to the current cloud. This process was repeated as long as additional genes could be included in the cloud. Finally, clouds within each other’s range, defined by the cluster_cloud_gap, were merged into one single large cluster.
In a final step, the statistical significance of the clouds was calculated. Two methods were available. One method used a binomial distribution in which the probability density was set to the number of dots divided by the area of the dot matrix. The other method took into account the removal of the tandem duplicates during a pre-processing step. Therefore, one dot per column and row was assumed to be present. For boxes larger than the tandem_gap, this assumption might be broken and the significance might be slightly overestimated. The binomial distribution supposed that one dot might be present at every position in the box. Hence, the second distribution would seemingly be a more realistic measure for the statistical significance of a cloud. As clouds could not be aligned, the profile search was automatically disabled with the cloud search.

For the human data set, the empirical FP rate was determined with a permutation test on several datasets, every time gradually increasing the cloud_gap from 5 to 55 and setting the cloud_cluster_gap to the cloud_gap plus five. We found that a cloud_gap of 15 (and thus a cloud_cluster_gap of 20) was the closest to the selected p-value and thus optimal for the human genome. Therefore, these settings were applied for further evaluation of the synteny mode. (Supplementary Table 3). Note that a wrong choice of parameters might cause an avalanche effect in which the bounding box keeps growing because new dots are found in the window frame.

On the human data set with the collinear search, 544 anchor point pairs were reported and the cloud search 2215 both, with a p-value of 0.01. Comparison of the number of block-duplicated genes revealed that the synteny mode detected nearly 4-fold more genes in significant syntenic blocks. Therefore; the synteny mode is recommended to detect highly diverged homologous regions, such as those derived from the WGD in vertebrates but also between species with a large evolutionary distance.
Supplementary Method 2: Parameter landscape for different eukaryotic model species

As the two user-defined parameters, the `gap_size` and `q_value`, have a direct impact on the accuracy and sensitivity of the collinearity detection, the optimal settings of these parameters depended on the properties of the data set. Within inter-species comparisons, closely related organisms would have a strongly dense diagonal in the GHM and, as the evolutionary distance increased, they would gradually break into smaller and sparser regions and, eventually, become undetectable. Intra-species properties are strongly linked with large-scale duplications. *Arabidopsis thaliana* has two relatively recent WGDs on top of an ancient hexaploidization shared with all other dicotyledonous plants (4,40,69) and, therefore, contains a mix of recent large-scale duplications and ancient homologous regions with abundant posterior gene loss and rearrangements. Intra-species GHMs for both human and yeast will result into a GHM with smaller, sparse and more diverged collinear regions. *Drosophila melanogaster* has no known WGD, neither ancient nor recent, and, thus, was excluded from further analysis.

To determine the effect of different parameters settings, the FP rate was monitored with a variety of setting combinations for the `gap_size` and `q_value` on several data sets. In this manner, the setting combination was determined in which the highest number of valid base clusters was reported while keeping the FP rate under control.

On denser genomes, due to recent duplications, such as *Arabidopsis* (70), the impact of the `q_value` remained rather small, but the `gap_size` became important to keep the FP rate under control, because the analytical `p-value` underestimated the probability to find
clusters with large gaps (Supplementary Table 4). In the human and yeast genomes, due to the sparse GHMs, a very low FP rate was observed and, thus, both the \textit{gap\_size} and \textit{q\_value} could be relaxed to find additional, highly diverged, duplicated regions (Supplementary Tables 5 and 6). As collinearity diverged more and more, gene order was no longer conserved and clusters no longer appeared as diagonal lines in the GHM, but as dense clouds. Therefore, application of the synteny mode is recommended for highly diverged genomes. Ideally, settings should be selected that maximize the amount of collinearity detected while keeping the FP rate close to the selected \textit{p\_value} threshold. For datasets with dense GHMs (species with recent WGDs and inter-species collinearity), a \textit{p\_value} threshold of 0.01 is advised in combination with a \textit{gap\_size} of 30 and a \textit{q\_value} of 0.75. As the average GHM density decreases, both parameters can be further relaxed to a \textit{gap\_size} of 55 and a \textit{q\_value} of 0.5.

### Supplementary Method 3: Impact of genome sequence assembly on the detection of collinearity

Like many comparative genomic studies, the detection of collinearity is highly dependant on the quality of the input data. Genomes sequenced to low coverage and/or using short reads usually are provided as a set of scaffolds rather than long pseudomolecules. The presence of such gaps in the genome sequence can manifest as false breaks in collinear regions, hence reducing the amount of collinearity found between two genomes.

To measure the impact of the overall genome quality is on the detection of collinearity, all available primate genomes as well as the treeshrew (\textit{Tupaia belangeri}) genome were
mapped onto the high-quality mouse genome. As all these species have a similar evolutionary distance to mouse and no extensive differences in generation time and lifestyle are known, a comparable fraction of genome collinearity can be expected with the mouse genome. The human (51), chimpanzee (*Pan troglodytes*) (52) and orangutan (*Pongo pygmaeus*) (71) genomes, all high-quality genomes, could be mapped to around 94% of the mouse genome (Supplementary Table 8). However, the gorilla (*Gorilla gorilla*) genome, sequenced to a 36x coverage using Solexa sequencing (72), mapped to only 8% of the mouse genome. The other genomes, even though there are all sequenced to approximately 2x coverage, showed a large variability (Supplementary Table 8). Clearly, the quality of the assembly, reflected in the number of scaffolds, rather than the sequencing depth accounts for most of the observed differences.
Supplementary Figure 1. Flowchart of the full i-ADHoRe 3.0 algorithm. Steps highlighted with green filled boxes can be executed in parallel.
Supplementary Figure 2. Correlation between empirical FP rate and the selected p-value cut-off. Inclusion of a correction for multiple-hypothesis testing (FDR and Bonferroni) results in an observed p-value closely reflecting the selected value. The recommended p-value range is $10^{-3}$ to $10^{-1}$. 
**Supplementary Figure 3.** Parallel speed-up in function of the number of processes used. The profile searches have a harder load to balance (smaller granularity) and therefore are not as efficient to run in parallel compared to the level 2 detection.
Supplementary Figure 4. Number of genes found in regions conserved in \( n \) species, as detected with Cyntenator, MCScan and i-ADHoRe. For a fair comparison with Cyntenator, only the counts conserved in all five species were considered.
Supplementary Figure 5. Conservation plot of the human chromosome 3. The height of the bars marks the number of species showing collinearity with that part of the chromosome. The areas in dark blue correspond to multiplicons significantly enriched (p-value < 0.05) for coexpressing or interacting gene pairs. The red line indicates the average conservation level whereas the horizontal gray bar gives the 5% of genes with the highest conservation level.
| Tool          | Type                | Comment                                                                 | Reference |
|--------------|---------------------|-------------------------------------------------------------------------|-----------|
| ADHoRe      | Pairwise – gene colinearity |                                                                          | (73)      |
| i-ADHoRe 2.0 NW | Multiple – gene colinearity | Needleman–Wunsch                                                         | (25)      |
| i-ADHoRe 2.0 GG | Multiple – gene colinearity | Greedy Graph                                                              | (25)      |
| LineUp      | Pairwise – gene colinearity | Two algorithms, both using a statistical model (probabilistic) one more precise but slow, based on markers (though these can be homologous genes) | (74)      |
| CloseUp     | Pairwise – shared gene density | Provides benchmark study incl. CloseUp, LineUp and ADHoRe                | (75)      |
| ColinearScan | Pairwise – gene colinearity |                                                                   | (76)      |
| FISH        | Pairwise – gene colinearity |                                                                           | (77)      |
| GeneSyn     | Pairwise – gene colinearity |                                                                           | (78)      |
| DAGchainer  | Pairwise – gene colinearity | DAG based                                                                | (79)      |
| OSfinder    | Pairwise – anchor colinearity | Provides benchmark study incl. ADHoRe and DAGchainer; stochastic Markov chain models | (80)      |
| SyMAP       | Pairwise – anchor synteny | BLAT to find anchors, makes DAG from anchors and scores using a distance function. | (81,82)  |
| MCScan      | Multiple – gene colinearity | Makes use of the transitivity property of collinearity.                  | (40)      |
| DiagHunter  | Pairwise - gene colinearity | Looks for significant diagonals. (Includes GenoPix2D to visualize output) | (83)      |
| Cinteny     | Pairwise – uses markers | Detects synteny blocks and calculates reversal distance                   | (84)      |
| Syntenator  | Multiple             | Uses Partial Order Graphs (POG)                                          | (44)      |
| Cyntenator  | Multiple             | Successor of Syntenator                                                  | (39)      |
| QUOTA-ALIGN | Pairwise             | Uses Binary Integer Programming , takes into account a user defined ploidy relations between the species of interest | (85)      |
| DRIMM       | Multiple             | Uses A-Bruijn graphs to generate synteny blocks                          | (86)      |

**Supplementary Table 1.** Comparison of different existing gene based colinearity detection tools. A more elaborate comparison between i-ADHoRe 3.0 and other tools that support multiple genomes (MCScan and Cyntenator) is included in the main paper.
| Species                  | Coverage        | #scaffolds | #protein coding genes | High quality subset |
|--------------------------|-----------------|------------|-----------------------|---------------------|
| Aedes aegypti            | hi-coverage     | 1612       | 15419                 | X                   |
| Anolis carolinensis      | 6.3x            | 2508       | 17660                 | X                   |
| Anopheles gambiae        | hi-coverage     | 6          | 12457                 | X                   |
| Bos taurus               | 7x              | 1133       | 21048                 | X                   |
| Caenorhabditis elegans   | hi-coverage     | 6          | 20176                 | X                   |
| Canis familiaris         | hi-coverage     | 41         | 19305                 | X                   |
| Cavia porcellus          | 6.79x           | 373        | 18673                 | X                   |
| Chloepus hoffmanni       | 2.05x           | 10959      | 12393                 |                     |
| Ciona intestinalis       | 11x             | 1628       | 14180                 | X                   |
| Ciona savignyi           | hi-coverage     | 343        | 11604                 | X                   |
| Danio rerio              | 6.5-7x          | 803        | 24147                 | X                   |
| Dasypus novemcinctus     | 2x              | 10408      | 14803                 |                     |
| Dipodomys ordii          | 1.85x           | 9840       | 15798                 |                     |
| Drosophila melanogaster  | hi-coverage     | 14         | 14141                 | X                   |
| Echinops telfairi        | 2x              | 12605      | 16562                 | X                   |
| Equus caballus           | 6.79x           | 105        | 20436                 | X                   |
| Erinaceus europaeus      | 1.86x           | 11710      | 14588                 |                     |
| Felis catus              | 1.87x           | 6718       | 15048                 |                     |
| Gallus gallus            | hi-coverage     | 52         | 16736                 | X                   |
| Gasterosteus aculeatus   | 11x             | 560        | 20787                 | X                   |
| Gorilla gorilla          | 35x             | 12076      | 16724                 |                     |
| Homo sapiens             | hi-coverage     | 84         | 21673                 | X                   |
| Loxodonta africana       | 7x              | 9590       | 15578                 | X                   |
| Macaca mulata            | 6.1x            | 750        | 21905                 | X                   |
| Microcebus murinus       | 1.93x           | 6929       | 16319                 |                     |
| Monodelphis domestica    | 7.33x           | 11         | 19466                 | X                   |
| Mus musculus             | hi-coverage     | 136        | 23497                 | X                   |
| Myotis lucifugus         | 1.7x            | 8772       | 16228                 |                     |
| Ochotona principes       | 1.93x           | 7406       | 15993                 |                     |
| Ornithorhynchus anatinus | 6x              | 8233       | 17951                 | X                   |
| Oryctolagus cuniculus    | hi-coverage     | 10573      | 15438                 | X                   |
| Oryzias latipes          | hi-coverage     | 886        | 19686                 | X                   |
| Otolemur garnettii       | 1.5x            | 7489       | 15443                 |                     |
| Pan troglodytes          | 6x              | 50         | 19829                 | X                   |
| Pongo pygmaeus           | 6x              | 52         | 20068                 | X                   |
| Procavia capensis        | 2.19x           | 10616      | 16044                 |                     |
| Pteropus vampyrus        | 2.63x           | 6385       | 16990                 |                     |
| Rattus Norvegicus        | hi-coverage     | 22         | 22503                 | X                   |
| Saccharomyces cerevisiae | hi-coverage     | 18         | 6698                  | X                   |
| Sorex araneus            | 1.9x            | 9911       | 13187                 |                     |
| Species                          | coverage       | #scaffolds | #protein coding genes |
|---------------------------------|----------------|------------|-----------------------|
| *Spermophilus tridecemlineatus* | 1.90x          | 8011       | 14827                 |
| *Taeniopygia guttata*           | 6x             | 68         | 17148                 |
| *Takifugu rubripes*             | hi-coverage    | 1930       | 18523                 |
| *Tarsius syrichta*              | 1.82X          | 11689      | 13615                 |
| *Tetraodon nigroviridis*        | hi-coverage    | 27         | 19602                 |
| *Tupaia belangeri*              | 2x             | 8248       | 15458                 |
| *Tursiops truncatus*            | 2.59x          | 5650       | 16537                 |
| *Vicugna pacus*                 | 2.51x          | 5004       | 11752                 |
| *Xenopus tropicalis*            | 7.65x          | 2543       | 18023                 |

Plaza version 1.0

Supplementary Table 2. Overview of all species included in the Ensembl and Plaza dataset.

Along with the coverage, the number of scaffolds and the number of coding genes. Genomes assembled into chromosomes are shown in bold.
| Cloud_gap Settings | FP-Rate     |
|-------------------|------------|
| 5                 | 4.27E-04   |
| 10                | 5.25E-03   |
| 15                | 2.14E-02   |
| 20                | 5.42E-02   |
| 25                | 1.06E-01   |
| 30                | 1.78E-01   |
| 35                | 2.77E-01   |
| 40                | 3.98E-01   |
| 45                | 5.22E-01   |
| 50                | 6.47E-01   |
| 55                | 7.90E-01   |

Supplementary Table 3. Empirical Estimated FP-Rates using the synteny mode on the human dataset (p-value cutoff 10^-2) In bold settings with a p-value near or better than the selected value are indicated.
Parameter combinations with a FP-rate equal or below the p-value cutoff specified (10^-2) are indicated in bold.

**Supplementary Table 4.** FP-rate on the basecluster level on the Arabidopsis dataset.

| gap | \(r^2\) 0,5 | \(r^2\) 0,6 | \(r^2\) 0,7 | \(r^2\) 0,8 | \(r^2\) 0,9 | \(r^2\) 1,0 |
|-----|-------------|-------------|-------------|-------------|-------------|-------------|
| 15  | 2,74E-04    | 2,74E-04    | 2,74E-04    | 2,75E-04    | 3,33E-04    | 0,00E+00    |
| 20  | 1,33E-03    | 1,34E-03    | 1,40E-03    | 1,39E-03    | 1,52E-03    | 0,00E+00    |
| 25  | 4,72E-03    | 4,70E-03    | 4,54E-03    | 4,58E-03    | 4,33E-03    | 0,00E+00    |
| 30  | 1,02E-02    | 1,00E-02    | 9,84E-03    | 9,21E-03    | 9,36E-03    | 0,00E+00    |
| 35  | 2,17E-02    | 2,10E-02    | 1,98E-02    | 1,86E-02    | 1,81E-02    | 0,00E+00    |
| 40  | 3,97E-02    | 3,76E-02    | 3,50E-02    | 3,22E-02    | 2,89E-02    | 0,00E+00    |
| 45  | 7,43E-02    | 7,00E-02    | 6,53E-02    | 5,92E-02    | 5,01E-02    | 0,00E+00    |
| 50  | 1,27E-01    | 1,18E-01    | 1,11E-01    | 9,94E-02    | 8,36E-02    | 0,00E+00    |
| 55  | 1,97E-01    | 1,81E-01    | 1,65E-01    | 1,46E-01    | 1,21E-01    | 0,00E+00    |

**Supplementary Table 5.** FP-rate on the basecluster level on the Human dataset.

| gap | \(r^2\) 0,5 | \(r^2\) 0,6 | \(r^2\) 0,7 | \(r^2\) 0,8 | \(r^2\) 0,9 | \(r^2\) 1,0 |
|-----|-------------|-------------|-------------|-------------|-------------|-------------|
| 15  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 20  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 25  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 30  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 35  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 40  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 45  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 50  | 4,08E-04    | 4,65E-04    | 2,44E-04    | 2,38E-04    | 2,63E-04    | 0,00E+00    |
| 55  | 7,55E-04    | 6,38E-04    | 4,65E-04    | 4,55E-04    | 5,13E-04    | 0,00E+00    |

**Supplementary Table 6.** FP-rate on the basecluster level on the yeast dataset.

| gap | \(r^2\) 0,5 | \(r^2\) 0,6 | \(r^2\) 0,7 | \(r^2\) 0,8 | \(r^2\) 0,9 | \(r^2\) 1,0 |
|-----|-------------|-------------|-------------|-------------|-------------|-------------|
| 15  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 20  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 25  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 30  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 35  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 40  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 45  | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    | 0,00E+00    |
| 50  | 1,48E-03    | 1,48E-03    | 1,48E-03    | 1,43E-03    | 1,07E-03    | 0,00E+00    |
| 55  | 1,72E-03    | 1,72E-03    | 1,38E-03    | 1,33E-03    | 1,00E-03    | 0,00E+00    |
**Supplementary Table 7.** See Excel file: Supplementary table 7.xlsx

| Genome       | Coverage          | # scaffolds | % genes contained in scaffolds of sufficient length | % of mouse genome covered |
|--------------|-------------------|-------------|-----------------------------------------------------|---------------------------|
| Human        | Finished          | 84          | 92,43%                                              | 94,34%                    |
| Chimp        | 6x                | 50          | 94,10%                                              | 93,89%                    |
| Gorilla      | 35x (solexa)     | 12076       | 54,14%                                              | 8,13%                     |
| Urangutang   | 6x                | 52          | 92,39%                                              | 94,16%                    |
| Tarsier      | 2x                | 11689       | 33,99%                                              | 0,80%                     |
| Mouse Lemur  | 2x                | 6929        | 65,51%                                              | 33,14%                    |
| Lemur        | 2x                | 7489        | 52,64%                                              | 21,49%                    |
| Tree Shrew   | 2x                | 8248        | 56,34%                                              | 18,57%                    |

**Supplementary Table 8.** Overview of the coverage of different genomes and the number of scaffolds/pseudomolecules the primate genomes are provided in. Additionally the percentage of genes that is present on scaffolds of sufficient length (>= the minimal number of homologous gene pairs required to detect collinearity between two segments) is indicated and the fraction of the mouse genome that shows significant collinearity with each genome.