Title: Optimization of capture protocols across species targeting up to 32000 genes and their extension to pooled DNA

Authors:

Cédric Mariac[1], Kévin Bethune[1], Sinara Oliveira de Aquino[1,2], Mohamed Abdelrahman[1,3], Adeline Barnaud[1], Claire Billot[4,5], Leila Zekraoui[1], Marie Couderc[1], Ndjido Kané[6], Alan Carvalho Andrade[7], Pierre Marraccini[1,8], Catherine Kiwuka[9,10], Laurence Albar[1,11], François Sabot[1], Valérie Poncet[1], Thomas LP Couvreur[1], Cécile Berthouly-Salazar[1,6], Yves Vigouroux[1]

[1] DIADE, Univ Montpellier, IRD, CIRAD, Montpellier, France

[2] Federal Univ of Lavras, Brazil

[3] Rice Research and Training Center, Field Crops Research Institute, Agricultural Research Center, Kafrelsheikh, Egypt

[4] AGAP Institut, Univ Montpellier, CIRAD, INRAE, InstitutAgro-SupAgro, Montpellier, France

[5] CIRAD, UMR AGAP Institut, F-34398 Montpellier, France

[6] LMI LAPSE, ISRA, IRD, Dakar, Senegal

[7] EMBRAPA Coffee-INOVACAFE, Lavras, Brazil

[8] CIRAD – UMR DIADE, F-34398 Montpellier, France

[9] NARO, Kampala, Uganda

[10] Wageningen University and Research Centre, Wageningen, Netherlands

[11] Present address: PHIM Plant Health Institute of Montpellier, Univ Montpellier, IRD, CIRAD, INRAE, Institut Agro, Montpellier, France

Yves Vigouroux[1] Institut de Recherche pour le Développement, Univ Montpellier, Unité Mixte de Recherche Diversité Adaptation et Développement des Plantes (UMR DIADE), Montpellier, France ;

Abstract:

Premise:

In-solution based capture is becoming a method of choice for sequencing targeted sequence.

Methods and results:

We assessed and optimized a capture protocol in 20 different species from 6 different plant genus using kits from 20,000 to 200,000 baits targeting from 300 to 32,000 genes. We evaluated both the effectiveness of the capture protocol and the fold enrichment in targeted sequences. We proposed a protocol with multiplexing up to 96 samples in a single hybridization and showed it was an efficient and cost-effective strategy. We also extended the use of capture to pools of 100 samples and proved the efficiency of the method to assess allele frequency. Using a set of various organisms with different genome sizes, we demonstrated a correlation between the percentage of on-target reads *vs*. the relative size of the targeted sequences.

Conclusion:

Altogether, we proposed methods, strategies, cost-efficient protocols and statistics to better evaluate and more effectively use hybridization capture.

INTRODUCTION

The reduced representation library approach (RRB) has become a widely used tool in molecular ecology and phylogeny. Some approaches are based on DNA restriction, e.g. RAd-seq (Andrews et al., 2016) or genotyping-by-sequencing (Elshire et al., 2011; GBS), in which the studied polymorphism is not targeted. In other approaches hybridization capture is used notably in phylogenetic studies (Mandel et al., 2014, Weitemier et al., 2014, Kollias et al., 2015, Nicholls et al., 2015, Stephens et al., 2015a, Stephens et al., 2015b, McCartney-Melstad et al., 2016, Portik et al., 2016, Couvreur et al., 2019) and to a lesser extend to diversity studies (Asan et al., 2011, Rosani et al., 2014, Kistler et al., 2015). Among these approaches, in-solution hybridization with DNA or RNA probes are the most frequent. Although they are common in human genetic studies (Asan et al., 2011), there are still underused in molecular ecology.

To assess the success of in-solution hybridization, the main parameter is "fold increase", *i.e.* to what extent a target sequence is enriched compared to the rest of the genome. This statistic is referred to here as x-fold enrichment. It measures how effective the enrichment is in terms of increasing the proportion of the target. Another parameter is the percentage of reads mapping on the target sequence (on-target reads, in contrast to off-target reads). This parameter is useful to assess the cost effectiveness of the experiment. Both statistics, x-fold enrichment and the percentage of on-target reads, are used interchangeably in the literature, whereas in fact, they assess two different aspects of the capture protocol. Such an enrichment protocol allows the analysis of many individuals, which is extremely useful in population genetics or phylogenetic studies. In addition, the combining of several individuals in a single hybridization may further reduce the cost. Lastly, pooled DNA sequencing is also increasingly used to assess allele frequency in a population (Pool-seq, Futschik and Schlötterer, 2010). In this approach, several individuals of the same population are sequenced together to estimate population allele frequency (Futschik and Schlötterer, 2010). The usefulness of the pool-seq approach combined with capture have not yet been fully evaluated yet, but is a promising avenue for future study.

Here, we used a set of 21 plant species with different genome sizes and target sequences to optimize in-solution capture approaches. X-fold enrichment and the percentage of on-target

reads were calculated to assess whether the approach is really both cost and experimentally effective. We show that high multiplexing for a single hybridization is both efficient and cost effective. A further extension of this approach was developed for pooled DNA sequencing. Finally, we demonstrated a relationship between the size of the target sequence relative to the size of the genome and the percentage of on-target reads, allowing better design and efficient prediction in any specific experiment.

MATERIALS AND METHODS

Plants materials.

A total of 21 species from four families belonging to 13 genera *(*Annonaceae: *Anaxagorea, Annickia, Anonidium, Greenwayodendron, Monanthotaxis, Monocarpia, Monodora, Neo-uvaria;* Arecaceae*: Podococcus;* Poaceae*: Cenchrus, Digitaria, Oryza;* Rubiaceae*: Coffea)* were used in this study (Table S1a). Fresh leaves, dried leaves and tissues collected in herbarium were used for total DNA extraction following a previously described protocol (Mariac et al., 2006). DNA was extracted from individual plants. In addition, for pearl millet variety PE5487, leaves from 100 plants were bulked before DNA extraction using a poolseq approach.

Locus target selection and bait design.

Seven enrichment kits were tested (Table S1b). Five of the seven kits were specific to the present study and only two have already been described (*Arecaceae*: Heyduck et al., 2016; *Annonaceae*: Couvreur et al., 2019). The total size of the targets to be captured varied from 204 kb to 12 Mb. RNA baits designed varied from 80bp to 120bp and from a 3X to 0.5X tiling (Table S1b). The targets were exonic sequences available either from transcriptome assemblies for *Cenchrus* and *Digitaria* (Sarah et al., 2016), or in the case of *Cenchrus americanum* (Varshney et al., 2017), *Oryza sativa* (Kawahara et al., 2013), and *Coffea canephora* (Denoeud et al., 2014), from fully annotated genomes. Biotinylated baits were designed and synthesized by Mycroarray (Ann Arbor, Michigan, USA). Mycroarray ran the set of baits through RepeatMasker v4,0 (Smit et al. 2013) over the closest available reference genome (Table S1b) to avoid designing baits that target repetitive sequences.

Library preparation and sequencing. Libraries were prepared according to the protocol of Mariac et al. (2018). Briefly, DNA samples were sheared to yield 400-bp fragments. DNA was then repaired and tagged using 6-bp barcodes to allow further multiplexing. Real-time PCR was performed to generate ready-to-load libraries. These libraries were either immediately sequenced for a shotgun genomic sequencing or enriched by capture according to the Myselect protocol (Mycroarray) before sequencing.

We tested different multiplexing levels for maximum possible cost reduction. From 1 to 96 equimolar libraries were multiplexed in a single DNA capture reaction. We then assessed the impact of multiplexing on enrichment efficiency.

Using the poolseq protocol, we built 100 individual libraries from 100 individual plants, and two libraries made by bulking exactly the same 100 plants. One (mock) library corresponded to an equimolar mix of the DNAs extracted from the 100 individuals, while the other library corresponded to a single extraction of DNA from the pooled leaves of the same 100 individuals.

A total input of 500 ng DNA was used per capture and hybridization was performed at 65 °C for 18 h including blocking oligonucleotides with six inosines at the barcode location to reduce unspecific hybridization. The immobilization and washing steps were conducted as recommended by the supplier. After probe-target hybridization (at 98 °C for 5 mins) the resulting enriched libraries were amplified using the KAPA Biosystem Real Time PCR Kit (KK8221) according to the supplier's recommendations. Paired-end sequencing was performed on an Illumina MiSeq (2x150 bp at CIRAD, Montpellier, France or on the Hiseq2000 platform (Genotoul, Toulouse).

Bioinformatics analysis: demultiplexing, data cleaning, mapping, SNP calling.

Demultiplexing based on 6-bp barcodes was performed using a Python script DEMULTADAPT (https://github.com/Maillol/demultadapt), using a 0-mismatch threshold. Adapters and low-quality bases were removed using CUTADAPT 1.8 (Martin, 2011) with the following options: quality cut-off = 20, minimum over-lap = 7 and minimum length = 35 (Table S1c). Reads with a mean quality lower than 30 were discarded using a PERL script (https://github.com/SouthGreenPlatform/arcad-hts/blob/master/scripts/arcad_hts_2_Filter_Fastq_On_Mean_Quality.pl). Mapping was performed using bwa mem 0.7.5a-r405 (Li et al., 2009) and the selected targets as the reference. GATK v3.3-0-g37228af UnifiedGenotyper (McKenna et al. 2010) was used for SNP calling. For the poolseq approach, we only kept SNPs with no missing data with a minimum coverage of 100 across the 100 individuals and at least 50 reads per bulk. Allele frequencies were extracted from the VCF files using VCFtools v 0.1.14 (Danecek et al., 2011). For the poolseq bulk, allele frequencies were calculated based on the count of the reference and alternate alleles using VCFtools v 0.1.14 (Danecek et al., 2011).

Estimation of enrichment efficiency. We first calculated the percentage of on-target reads, *i.e.* the number of reads mapped on the target references divided by the total number of reads. We then calculated the x-fold enrichment (x-fold), *i.e.* the number of on-target reads after enrichment divided by the number of on-target reads without enrichment. The latter number was estimated based on sequencing of the whole genome (Table S1d). These two statistics were calculated for each library; averages and standard deviations were calculated per species and per multiplex (Table S1d).

Literature review. To gain comprehensive insight into the methodology, we compared our results with those obtained in previously published studies. Articles that described the use of a similar method on fresh and ancient DNA of plants and animals were downloaded and analyzed (see list of publications in supplementary file). For each study, when possible, we re-encoded or calculated both statistics, *i.e*. the x-fold enrichment and the percentage of on-target reads. For each of these studies, we also calculated the ratio of the total size of the targeted sequences to the total genome size. When necessary, the genome size was evaluated based on the litterature C-value (Dolezel et al., 2003, detailed in table S2).

RESULTS AND DISCUSSION

Results of multiplexing did not change either the x-fold or the number of on-target reads

We first analyzed the results obtained on three species in greater detail (Table S1d): *Cenchrus americanus*/*Pennisetum glaucum* (kit MIL-328), *Digitaria exilis* (Kit Fonio) and *Coffea canephora (*kit Coffee*)*.

For pearl millet, with a capture approach targeting 328 genes (MIL-328), the percentages of on-target reads after enrichment were 14.3 % (se=3.34), 17.3 % (se=0.21) and 18.4 % (se=0.16) for 1, 8 and 30 libraries per capture, respectively (Figure S1, Table S1d). A slightly higher number of on-target reads was retrieved with higher multiplexing (Kruskal-Wallis, H=12.38, p=0.002). The x-fold enrichment associated with multiplexing varied from 78.9, 99.3 and 106.6 for 1, 8 and 30 libraries per capture, respectively (Figure S1). Enrichment was more efficient with higher multiplexing (Kruskal-Wallis, H=17.82, p<0.0001).

For *Digitaria exilis*, with a capture targeting 3000 genes, the average percentages of on-target reads after enrichment were 83.5 % (se=0.009), 83.7 % (se=:0.851) and 85.3 % (se=1.55) for 1, 8 and 39 libraries per capture, respectively. The percentage of useful reads was slightly better with higher multiplexing (Kruskal-Wallis, H=8.19, p=0.016). The average x-fold values were similar (Kruskal-Wallis, H=4.01, p=0.134) with 33.6, 35.7 and 34.5 for 1, 8 and 30 libraries per capture, respectively (Figure S1).

For *Coffea canephora*, the targeted 323 genes (de Aquino et al. 2021) represented a total length of 1.3 Mb *i.e.* 0.2% of the whole genome (1C = 710 Mb). The percentage of reads on-target after enrichment of 70%. The x-fold level of target enrichment was 270 compared to a non-enriched genome sequencing, and only slightly lower than a single captured library x-fold enrichment (320).

Our results showed that the percentage of on-target reads was generally not affected by the number of individuals multiplexed for hybridization across the other kits and species (Figure S2). The percentage ranged from 15% for a single capture on pearl millet (MIL-358) to 80% on rice (Figure 1, Figure S2). The x-fold enrichment ranged from 5 on pearl millet to 400 for *Annonaceae* (Figure S2). We tested up to 96 individuals captured at once (Figure 1, Figure S2). Based on an estimated cost of USD3,600 for 16 captures (https://arborbiosci.com/products/custom-target-capture/), the capture price per sample amounted to USD225 without multiplexing, and to only to USD2,34 with multiplexing of 96 samples, as proposed in our protocol, which makes the cost really affordable. Our result thus

showed that multiplexing several individuals in a single capture hybridization is an efficient and cost-effective strategy.

Analysis of the poolseq approach with capture

To evaluate the accuracy of the capture approach for poolseq DNA, we compared the allele frequencies obtained from libraries performed on 100 individual DNA samples, on a bulk of 100 DNA samples (bulk 1), and on a DNA sample extracted from the pooled leaves of 100 individuals (bulk 2). We identified a total of 62,320 SNPs in the 100 individuals and two bulk samples. The allele frequencies between these two bulk samples (r^2=0.99, $p<10^{-20}$) were highly correlated, as well as between individual libraries and the bulks (r^2=0.98, $p<10^{-20}$ between individual libraries and bulk 2) (Figure 2). Capture thus made it possible to effectively retrieve the allele frequency of a bulk of individuals. Our protocol makes this approach more broadly accessible, even for large genomes. Such an approach could also be used for bulk segregant analysis (Takagi et al., 2013) to identify functional variations linked to specific traits.

Literature review and meta-analyses.

We retrieved or calculated the percentage of on-target reads from 22 studies (Table S2). Using these data and the seven studies presented here, we evidenced a logarithmic relationship (r^2=0.54, $p<2.10^{-5}$) between the percentage of on-target reads and the relative size of the targeted sequence (Figure 3). For small targets representing less than 1% of the genome, it is not uncommon to find a relatively low percentage of on-target reads (Figure 3), but the percentage of on-target reads increases exponentially up to almost 100% with an increase in the relative size of the target. This relationship makes it possible to predict the percentage of on-target reads based on the size of the genome and the size of the targeted sequence, meaning a preliminary assessment of the cost effectiveness of the experiment can be performed before design. An extreme example of capture design that targets a very short sequence (~500bp) will lead to very low on-target reads percentage, with 1% or less (Maggia et al., 2017). Recent results show that the protocol and two rounds of capture can be performed to increase on-target reads on target up to 70% (Mariac et al., 2018). With the low cost of capture for high multiplexing, two rounds of capture would be very cost effective.

Only 5 out of 22 (23%) studies in the literature provided the x-fold enrichment level. It was thus rather difficult to perform meta-analysis on this statistic. X-fold enrichment really measures the effectiveness of the experiment because it assesses how many more sequences are obtained after capture compared to a non-enrichment baseline. We suggest that both the percentage of on-target reads and x-fold enrichment should be more frequently reported in the literature. When the x-fold enrichment is not reported, it is difficult to really assess if and how well the capture experiment worked.

CONCLUSION:

We assessed capture protocols in different plant species, and showed high multiplexing per capture is an efficient strategy across species. We demonstrated that this capture strategy can easily be extended to pooled DNA samples to assess allele frequency. We highlighted a relationship that allowed the percentage of on-target reads to be estimated. The early estimation of this important statistic could guide the protocol and the size of the target sequence for more efficient use of these approaches. Finally, we highlighted the importance of reporting both the percentage of on-target reads and the x-fold enrichment in any capture experiment to allow better assessment of the approach.

ACKNOWLEDGEMENT

This work was funded by an ANR grant to YV; CBS, CM, YV, FS and by Agropolis Foundation (Reference ID 1402-003) through the « Investissements d'avenir » program (Labex Agro:ANR-10-LABX-0001-01). VP, SOA, CK, AA and PM were supported by two Agropolis Foundation - CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) projects, i.e. ID 1002-009 and ID 1402-003 (CLIMCOFFEA), through the Investissements d'avenir program (Labex Agro:ANR-10-LABX-0001-01), in the framework of I-SITE MUSE (ANR-16-IDEX-0006). MA had a fellowship from the French Embassy in Egypt, Institut Français d'Egypte and the Science and Technology Development Fund. The authors acknowledge the ISO 9001 certified IRD itrop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URLs: https://bioinfo.ird.fr/ and http://www.southgreen.fr. We thanks Anaïs Dequincey and Coralie Picard for help in developing some of the experiments.

BIBLIOGRAPHY

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., and Hohenlohe, P.A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet *17*, 81–92.

Asan, Xu, Y., Jiang, H., Tyler-Smith, C., Xue, Y., Jiang, T., Wang, J., Wu, M., Liu, X., Tian, G., et al. (2011). Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biol *12*, R95.

Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R., and Moritz, C. (2013). Unlocking the vault: next-generation museum population genomics. Mol Ecol *22*, 6018–6032.

Couvreur, T.L.P., Helmstetter, A.J., Koenen, E.J.M., Bethune, K., Brandão, R.D., Little, S.A., Sauquet, H., and Erkens, R.H.J. (2019). Phylogenomics of the major tropical plant family annonaceae using targeted enrichment of nuclear genes. Front Plant Sci *9*, 1941.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

de Aquino, S. O., Kiwuka, C., Tournebize, R., Gain, C., Marraccini, P., Mariac, C., Bethune, K., Couderc, M., Cubry, P., Andrade, A. C, Lepelley, M., Darracq, O., Crouzillat, D., Anten, N., Musoli, P., Manel, S.,

Vigouroux, Y., de Kochko, A., François, O., and Poncet, V. (2021). Adaptive potential of *Coffea canephora* from Uganda in response to climate change. Molecular Ecology, submitted.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science *345*, 1181–1184.

Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. Cytometry A *51*, 127–128; author reply 129.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE *6*, e19379.

Faircloth, B.C., Branstetter, M.G., White, N.D., and Brady, S.G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol Ecol Resour *15*, 489–501.

Folk, R.A., Mandel, J.R., and Freudenstein, J.V. (2015). A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: A phylogenomic example from *Heuchera* (Saxifragaceae). Appl Plant Sci *3*, apps.1500039.

Futschik, A., and Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics *186*, 207–218.

Guschanski, K., Krause, J., Sawyer, S., Valente, L.M., Bailey, S., Finstermeier, K., Sabin, R., Gilissen, E., Sonet, G., Nagy, Z.T., et al. (2013). Next-generation museomics disentangles one of the largest primate radiations. Syst Biol *62*, 539–554.

Hawkins, M.T.R., Hofman, C.A., Callicrate, T., McDonough, M.M., Tsuchiya, M.T.N., Gutiérrez, E.E., Helgen, K.M., and Maldonado, J.E. (2016). In-solution hybridization for mammalian mitogenome enrichment: pros, cons and challenges associated with multiplexing degraded DNA. Mol Ecol Resour *16*, 1173–1188.

Heyduk, K., Trapnell, D.W., Barrett, C.F., and Leebens-Mack, J. (2016). Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. Biol J Linn Soc *117*, 106–120.

Hugall, A.F., O'Hara, T.D., Hunjan, S., Nilsen, R., and Moussalli, A. (2016). An exon-capture system for the entire class Ophiuroidea. Mol Biol Evol *33*, 281–294.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice *6*, 4.

King, R., Bird, N., Ramirez-Gonzalez, R., Coghill, J.A., Patil, A., Hassani-Pak, K., Uauy, C., and Phillips, A.L. (2015). Mutation scanning in wheat by exon capture and next-generation sequencing. PLoS ONE *10*, e0137549.

Kirillova, I.V., Zanina, O.G., Chernova, O.F., Lapteva, E.G., Trofimova, S.S., Lebedev, V.S., Tiunov, A.V., Soares, A.E.R., Shidlovskiy, F.K., and Shapiro, B. (2015). An ancient bison from the mouth of the Rauchua River (Chukotka, Russia). Quaternary Res *84*, 232–245.

Kistler, L., Montenegro, Á., Smith, B.D., Gifford, J.A., Green, R.E., Newsom, L.A., and Shapiro, B. (2014). Transoceanic drift and the domestication of African bottle gourds in the Americas. Proc Natl Acad Sci USA *111*, 2937–2941.

Kistler, L., Ratan, A., Godfrey, L.R., Crowley, B.E., Hughes, C.E., Lei, R., Cui, Y., Wood, M.L., Muldoon, K.M., Andriamialison, H., et al. (2015a). Comparative and population mitogenomic analyses of Madagascar's extinct, giant "subfossil" lemurs. J Hum Evol *79*, 45–54.

Kistler, L., Newsom, L.A., Ryan, T.M., Clarke, A.C., Smith, B.D., and Perry, G.H. (2015b). Gourds and squashes (*Cucurbita* spp.) adapted to megafaunal extinction and ecological anachronism through domestication. Proc Natl Acad Sci USA *112*, 15107–15112.

Kollias, S., Poortvliet, M., Smolina, I., and Hoarau, G. (2015). Low cost sequencing of mitogenomes from museum samples using baits capture and Ion Torrent. Conservation Genet Resour *7*, 345–348.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). Thesequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Maggia, M.E., Vigouroux, Y., Renno, J.F., Duponchelle, F., Desmarais, E., Nunez, J., García-Dávila, C., Carvajal-Vallejos, F.M., Paradis, E., Martin, J.F., et al. (2017). DNA Metabarcoding of Amazonian Ichthyoplankton Swarms. PLoS ONE *12*, e0170009.

Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H., and Burke, J.M. (2014). A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. Appl Plant Sci *2*, apps.1300085.

Mariac, C., Luong, V., Kapran, I., Mamadou, A., Sagnard, F., Deu, M., Chantereau, J., Gerard, B., Ndjeunga, J., Bezançon, G., et al. (2006). Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L.] R. Br.) in Niger assessed by microsatellite markers. Theor Appl Genet *114*, 49–58.

Mariac, C., Vigouroux, Y., Duponchelle, F., García-Dávila, C., Nunez, J., Desmarais, E., and Renno, J.F. (2018). Metabarcoding by capture using a single COI probe (MCSP) to identify and quantify fish species in ichthyoplankton swarms. PLoS ONE *13*, e0202976.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J *17*, 10–12.

McCartney-Melstad, E., Mount, G.G., and Shaffer, H.B. (2016). Exon capture optimization in amphibians with large genomes. Mol Ecol Resour *16*, 1084–1094.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297–1303.

Nicholls, J.A., Pennington, R.T., Koenen, E.J., Hughes, C.E., Hearn, J., Bunnefeld, L., Dexter, K.G., Stone, G.N., and Kidner, C.A. (2015). Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). Front Plant Sci *6*, 710

Portik, D.M., Smith, L.L., and Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). Mol Ecol Resour *16*, 1069–1083.

Rosani, U., Domeneghetti, S., Pallavicini, A., and Venier, P. (2014). Target capture and massive sequencing of genes transcribed in *Mytilus galloprovincialis*. BioMed Res Intern *2014*, Article ID 538549.

Sarah, G., Homa, F., Pointet, S., Contreras, S., Sabot, F., Nabholz, B., Santoni, S., Sauné, L., Ardisson, M., Chantret, N., et al. (2016). A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Mol Ecol Resour *17*, 565-580

Schweizer, R.M., Robinson, J., Harrigan, R., Silva, P., Galverni, M., Musiani, M., Green, R.E., Novembre, J., and Wayne, R.K. (2016). Targeted capture and resequencing of 1040 genes reveal environmentally driven functional variation in grey wolves. Mol Ecol *25*, 357–379.

Smit, A.F.A., Hubley R., and Green, P. (2013-2015) *RepeatMasker Open-4.0*. 2013-2015 <http://www.repeatmasker.org>.

Stephens, J.D., Rogers, W.L., Mason, C.M., Donovan, L.A., and Malmberg, R.L. (2015a). Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. Am J Bot *102*, 910-920

Stephens, J.D., Rogers, W.L., Heyduk, K., Cruse-Sanders, J.M., Determann, R.O., Glenn, T.C., and Malmberg, R.L. (2015b). Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. Mol Phylogenet Evol *85*, 76–87.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., et al. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J *74*, 174–183.

Varshney, R.K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., Zhang, H., Zhao, Y., Wang, X., Rathore, A., et al. (2017). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. Nat Biotechnol *35*, 969–976.

Weitemier, K., Straub, S.C.K., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A., and Liston, A. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. Appl Plant Sci *2*, apps.1400042.

Figure 1. Percentage of on-target reads across species and capture design.

Percentage of on-target reads (grey bars) compared to unenriched libraries (dark bars) and their confidence intervals. For each. the species studied are *Digitaria exilis* (fonio), *Cenchrus americanun* (with three different kit: MIL-328, MIL-CLR, MIL-EXOME, see Table SXX), *Oryza* spp. (Rice), *Annonacea* spp. (Annonaceae), *Coffea* spp. (Coffea), several palm species (Palms).

Figure 2. Correlation of expected and observed allele frequencies using poolseq capture protocols.

Calculated correlations between allele frequencies estimated based on individual capture of 100 individuals (Freq_B1), and on capture of bulk samples. One bulk sample was made of an equimolar concentration of DNA of 100 individuals (freq_mock), the other was made of a DNA extract of 100 pooled pieces of leaves from the same 100 individuals (freq_B32). The correlations were highly significant between all the three experimental conditions ($r^2 \geq 0.98$, $p < 10^{-20}$).

Figure 3. Relationship between the percentage of on-target reads and the relative size of the targeted sequence

Significant relationship between the logarithm of the relative targeted sequence (size of the target divided by the size of the genome) and the number of on-target reads ($r^2=0.54$, $p<2\ 10^{-5}$). Red dots represent data collected in the present study and blue dots represent data retrieved from the literature (Table S2).

Figure 1.

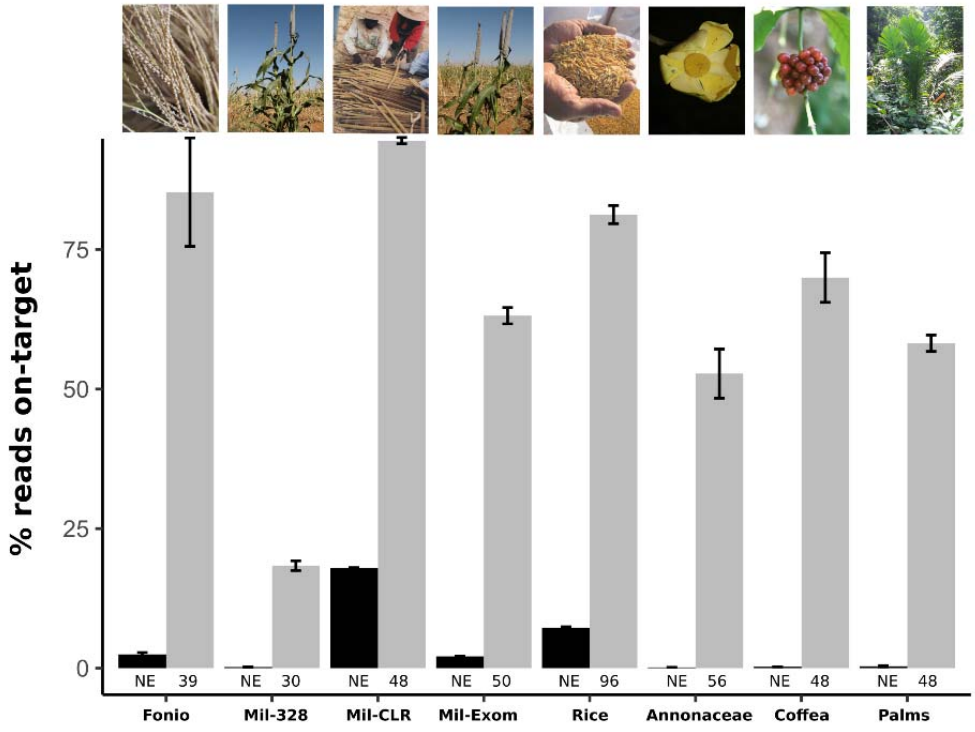


Figure 2.

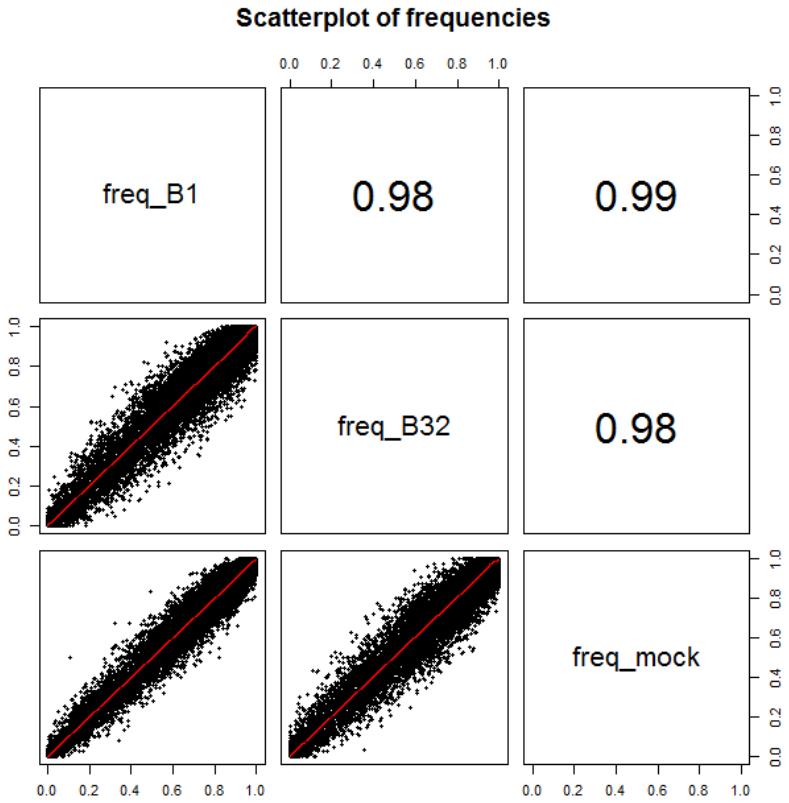
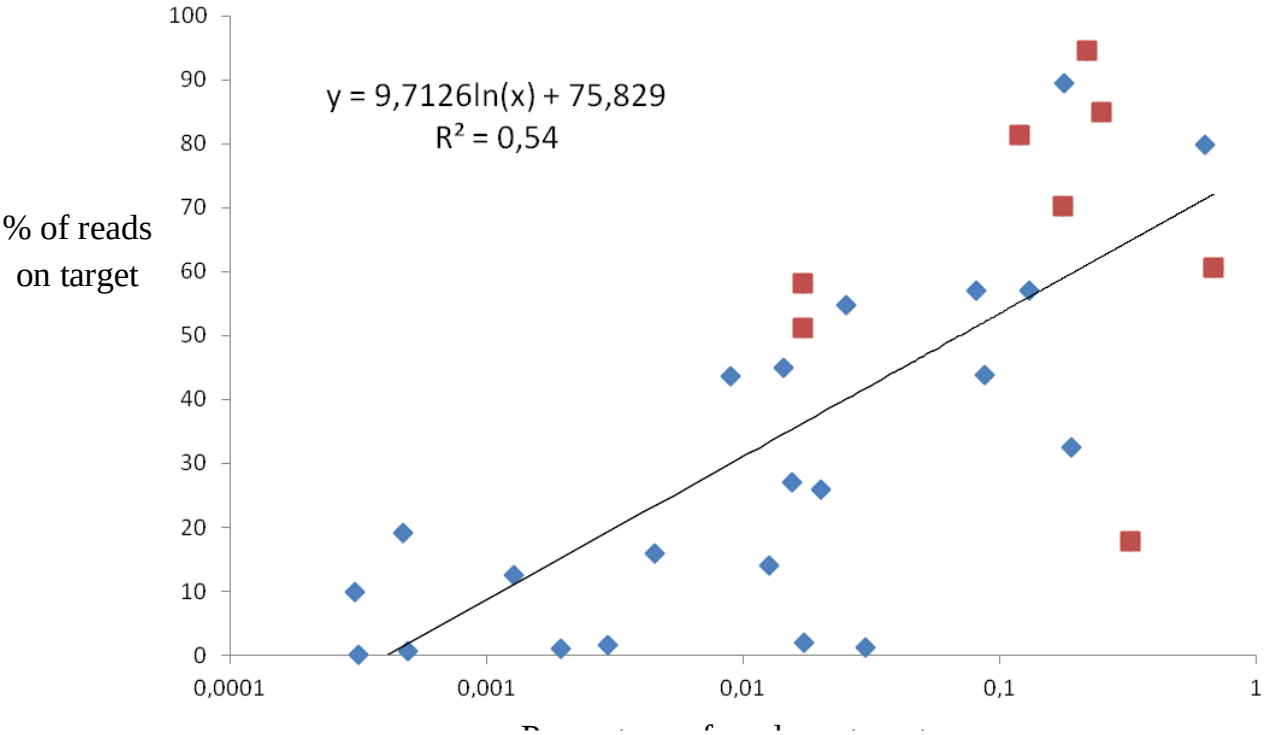


Figure 3.



Data availability :

Sequencing reads were deposited in the National Center for Biotechnology Information

Sequence Read Archive (BioProject ID: PRJNA431698, BioSample accessions:

SAMN08625127- SAMN08625640).

Supporting information

Fasta files of baits and target used for each kit:

TARGET-MIL328.fasta
TARGET-MIL-CLR.fasta
TARGET-palms.exons.final.fasta
TARGET-RICE.fasta
bait-Annonaceae_nuc.fas
baits- MIL-EXOME-bait-80-160-first7-fixed.fas
baits-COFFEE.fas
baits-MIL-328.fas
baits-MIL-CLR.fas
baits-PALM_EXONS.fas
baits-RICE.fas
TARGET-MIL-EXOME-152169_stringent_baits-coordv1.1.bed
TARGET-Annonaceae_nuc.fasta
TARGET-COFFEE.fasta
TARGET-FONIO.fasta

Supplementary figures

Figure S1. X-fold enrichment and percentage of on-target reads for *Pennisetum glaucum* and *Digitaria exilis*

We tested different multiplexing of samples for a single capture experiment (1, 8 and 39 libraries per capture for *Pennisetum glaucum*; 1, 8 and 39 libraries per capture for *Digitaria exilis*). Experiments included non-enriched libraries as controls. Both the fold enrichments and the percentages of on-target reads obtained with different multiplexing were very similar and even tended to be slightly better with an increase in the number of samples.

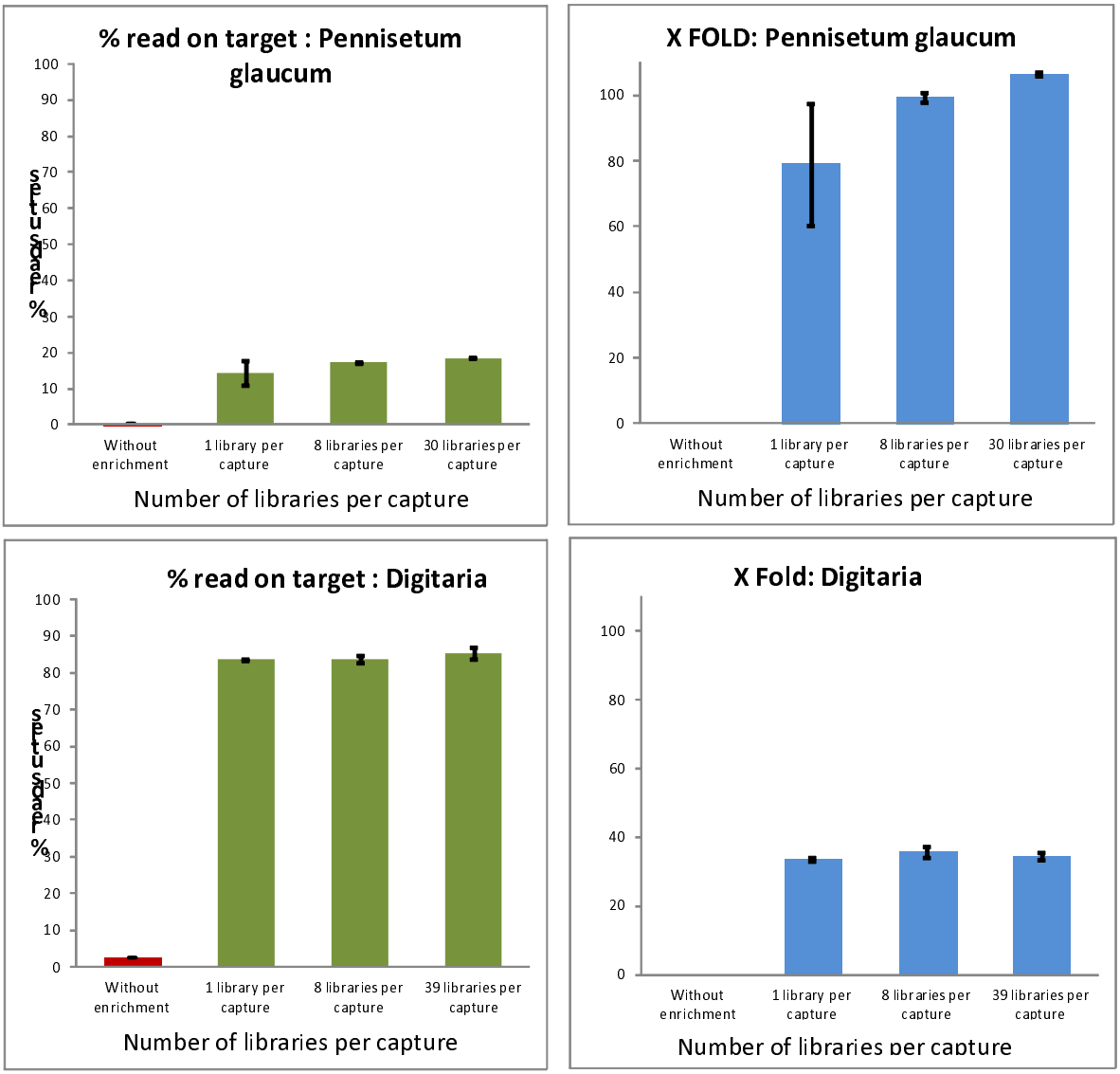


Figure S2. X-fold and percentage of on-target reads for all the species

For all the species (7 different capture kits), we tested different multiplexing of samples for a single capture experiment. The experiment included non-enriched DNA sequence as a control.

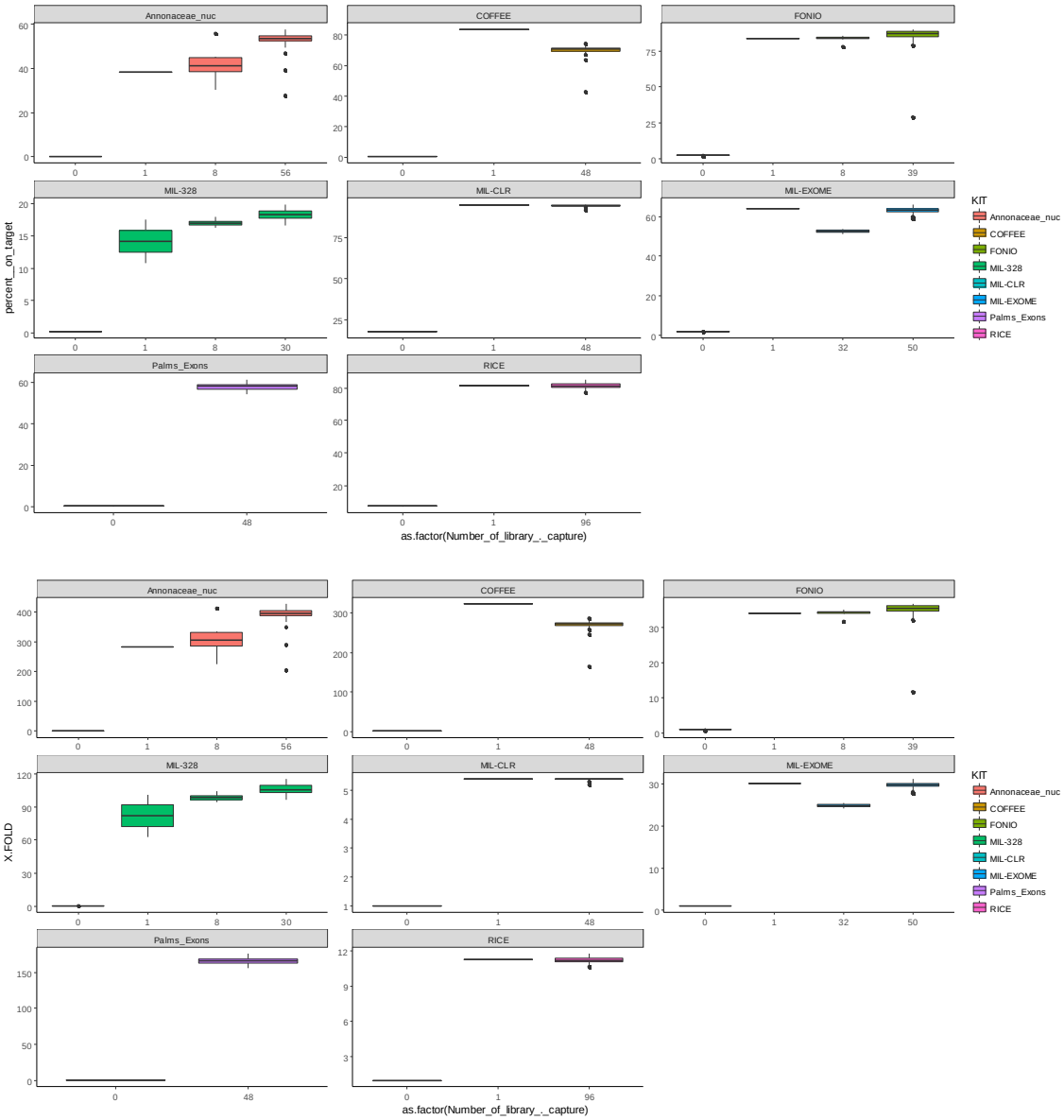


Table S1. Passport data, bait design, bioinformatics and raw mapping data.

Table S1a. passport data of all samples used in this study.

The table list : plant name, the collection ID of the accession, the species name, type, origin of the material and the country where the accession was sampled.

TableS1b. Details of the kit used in this study.

Table S1c. Option used for bioinformatic analysis and mapping reads

Table S1d. List of individual and each sequencing files

For each individual, its accession name (Accession_IR, Accession_ID2), the species, the tag used for sequencing (internal information RUN-index-TAG), an experimental code (Code), the library preparation (capture or direct shotgun without capture) and the kit used for the experiment are given.
We also give the number of libraries used for a single capture experiment, the number of Illumina reads, the number of reads mapping on target, the percentage of read on target, the x-fold enrichment
We also give reference of each sample in GenBank (Biosample identification, link and name of the fastq files)

see excel files

Table S2. List of studies in which genome size, capture target and the percentage of useful reads are included.

	Target size	Genome size	ratio	% of useful reads	x-fold	Reference
Animal	15500	4900000000*	0,000316	0,0093		Kirillova et al., 2015
Animal	15500	3160000000*	0,000491	0,70	98,5	Kistler et al., 2015a
Plant	107640	5530000000*	0,00195	1.00		Stephens et al., 2015a
Plant	150000	500000000*	0,03	1,26		Kistler et al., 2015b
Plant	128110	4350000000*	0,00295	1,62		Stephens et al., 2015b
Plant	86000	500000000*	0,0172	2.00		Kistler et al., 2014
Animal	15500	5080000000*	0,000305	9.88	1 to 2	Kollias et al., 2015
Insect	15600	1222000000	0,00128	12.50	178 to 744	Faircloth et al., 2015
Animal	2530000	2E+10*	0,0127	14.00		McCartney-Melstad et al., 2016
Plant	102000	2250000000	0,00453	15.95		Heyduk et al., 2016
Animal	15000	3170000000*	0,00047	19.20	15 to 2700	Hawkins et al., 2016
Plant	2000000	1E+10	0,02	26.00		King et al., 2015
Plant	580680	3740000000*	0,0155	27.00		Mandel et al., 2014
Plant	1600000	840000000*	0,19	32.50		Weitemier et al., 2014
Animal	4000000	4600000000	0,0870	43.75		Portik et al., 2016
Animal	285165	2000000000	0,014	45.00		Hugall et al., 2016
Plant	378553	1500000000*	0,025	54.70		Folk et al., 2015
Animal	1500000	1160000000*	0,129	57.00	50	Rosani et al., 2014
Plant	259313	41000000*	0,63	79.90		Nicholls et al., 2015
Animal	5000000	2810000000*	0,178	89.47		Schweizer et al., 2016

We only consider studies for which genome size, capture target and the percentage of useful reads are included or could be easily calculated. For each study, whether the analysis is based on a plant, animal or insect, the target size of the baits (bp), the genome size (bp), the ratio of target divided by genome size (ratio), the percentage of useful read *i.e.* percentage of reads on target, the x-fold enrichment (x-fold) and the article reference are given.

* Genome size based on the C-Value.