Chemo- versus immuno-precipitation of G-quadruplex-DNA (G4DNA): a direct comparison of the efficiency of the antibody BG4 versus the small-molecule ligands TASQs for G4 affinity capture

Yilong Feng, Zhenyu Luo, Francesco Rota Sperti, Ibai Valverde, Wenli Zhang, David Monchaud

Chemo- *versus* immuno-precipitation of G-quadruplex-DNA (G4-DNA): a direct comparison of the efficiency of the antibody BG4 *versus* the small-molecule ligands TASQs for G4 affinity capture

Yilong Feng,[1,#] Zhenyu Luo,[1,#] Francesco Rota Sperti,[2]
Ibai E. Valverde,[2] Wenli Zhang[1,⋆] and David Monchaud[2,⋆]

[1]State Key Laboratory for Crop Genetics and Germplasm Enhancement, CIC-MCP, Nanjing Agricultural University, Nanjing, P.R. China. [2]Institut de Chimie Moléculaire, ICMUB CNRS UMR 6302, UBFC Dijon, France.

[#]These authors contributed equally to this work. [⋆]E-mail: wzhang25@njau.edu.cn, david.monchaud@cnrs.fr

Abstract. The search for genomic G-quadruplex (G4) motifs is motivated by their involvement in key cellular processes (*e.g.*, transcription, replication) and corresponding dysregulations underlying genetic diseases. Sequencing-based methods have been developed to assess G4 DNA formation genome-wide, including G4-seq that detects G4s in purified DNA (so called *in vitro*) from human lymphocytes using the G4 stabilizer PDS, and G4 ChIP-seq that detects G4s in the chromatin of human keratinocytes (so called *in vivo*) using the G4-specific antibody BG4. We recently reported on the use of a small molecule, BioTASQ, as a proxy for BG4 for assessing the transcriptomic prevalence of G4-RNA *in vivo* using G4RP-seq. Here, we provide the first direct and unbiased comparison of the G4 capturing ability of small molecules BioTASQ and its new derivative BioCyTASQ *versus* the antibody BG4, developing G4DP-seq which makes the genome-wide detection of G4s *in vitro* and *in vivo* more straightforward, reproducible and reliable.

Keywords: G-quadruplex; BG4; IP-seq ; TASQ; G4DP-seq

Introduction

G-quadruplex-DNA (G4-DNA, or G4) are DNA structural motifs originating in the folding of guanine (G)-rich DNA sequences, when freed from the duplex constraint, into a four-stranded structure.[1, 2] The stability of these motifs is provided by both the self-assembly of Gs to form G-quartets and the self-stacking of several contiguous G-quartets (**Figure 1A**).[3] Gs are thus not randomly distributed along the G4-forming sequences (G4FS) but gathered into G-runs usually comprising 2 to 4 Gs. The repetitive nature of these sequences makes them readily detectable genome-wide: the bioinformatics processing of the human genome in the search for the general sequence $G_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}N_xG_{\geq 3}$ (where N is any intervening nucleobase, x ranging from 1 to 7) led to the identification of >300,000 putative G4FS.[4, 5] The length of the connecting loops (the x value) was then extended to ≤ 15, ≤ 21 and then ≤ 25, which mechanically increased the number of G4FS to >1,000,000.[6-9]

This high G4 density prompted researchers to demonstrate their existence *in vitro*. To this end, the G4-seq technique[6] was developed and applied to purified single-stranded DNA, using G4-promoting conditions (*i.e.*, in a buffer with a high K^+-content or in the presence of the G4-stabilizing small-molecule pyridostatin,[10] PDS). The folded and stable G4s were then detected through a polymerase stop assay, in which the stabilized G4 motifs pause the polymerase procession, which creates an erroneous incorporation of nucleotides in the downstream sequence. The resulting, highly mismatched sequences are then easily visualizable through a Phred quality score analysis, which led to the straightforward detection of the polymerase stop sites, that is, the G4 sites. G4-seq identified >500,000 G4FS in K^+-rich buffers, and >700,000 G4FS in presence of PDS within the human genome, and was subsequently applied to 12 different species (including human, mouse, bacteria and *Arabidopsis*, a model plant species)[11] to track and highlight the diversity of abundance and location of G4FS (being particularly prevalent in mammals).

This high G4 density prompted researchers to investigate the existence of G4s *in cella*: to this end, an antibody-based G4-immunoprecipitation protocol was developed, the G4 ChIP-seq,[12, 13] using fragmented chromatin from fixed human keratinocyte cells from which folded G4s were precipitated using BG4,[14] a G4-specific BG4 antibody (Ab, **Figure 1B**). G4 ChIP-seq identified *ca*. 10,000 G4s, which represent a minor fraction (2%) of G4FS detected by G4-seq, likely due to the repressive role of the chromatin packing. We followed a convergent approach to assess the prevalence of G4s in the genome of plants (rice). This technique, termed BG4-IP-

seq,[15] was implemented with purified and fragmented rice genomic DNA in G4-promoting conditions (*i.e.*, in a buffer with a high K^+-content) from which folded G4s were precipitated using BG4, as above. BG4-IP-seq identified *ca*. 20,000 G4FS, which represents a minor fraction (5%) of *in silico* predicted G4FS, but far higher than the number of G4FS detected by G4-seq in another model plant species, the *Arabidopsis* (*ca*. 2,000 G4FS), as a result of the difference in the implemented technique.

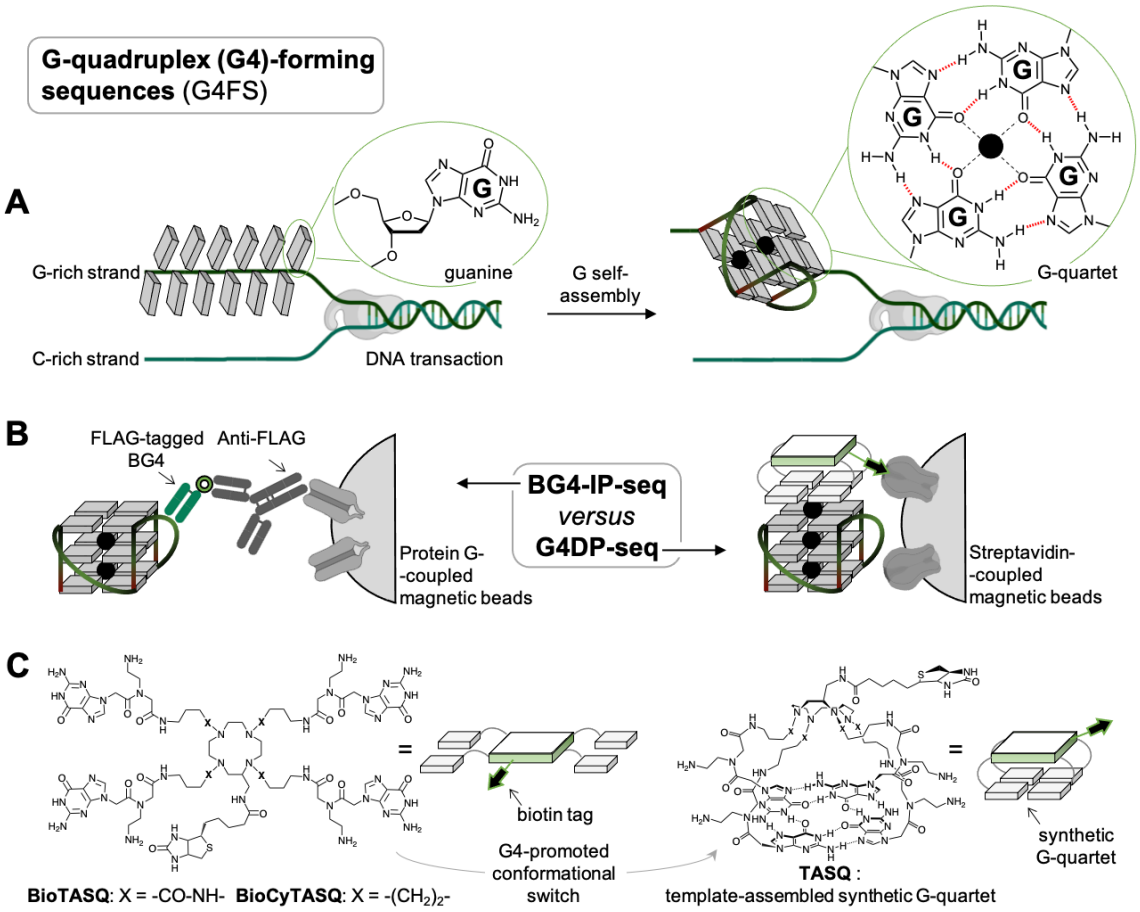

Figure 1. Schematic representations of a G-quadruplex (G4)-folding from a guanine (G)-rich DNA sequences (**A**) and of the two affinity capture techniques studied here (**B**), the BG4-IP-seq performed with the antibody BG4 and the G4DP-seq performed with two biotinylated small molecules, BioTASQ and BioCyTASQ (**C**).

The reliability of the BG4-IP-seq protocol, along with the straightforward access to rice genomic DNA, makes this system ideal for comparing the pull-down efficiency of the antibody BG4 *versus* the small-molecular baits BioTASQ[16, 17] and its new derivative BioCyTASQ[18] (**Figure 1C**). These two TASQs (for template-assembled synthetic G-quartets)[19] are biotinylated biomimetic ligands that were successfully employed for fishing G4-RNAs out from human cell lysates, in a protocol referred to as G4RP-seq (for G4-RNA precipitation and sequencing),[16, 20]

which belong to the toolbox of seq-based techniques used to interrogate G4-RNA prevalence,[21] comprising rG4-seq,[22] DMS-seq,[23] Keth-seq[24] and SHALiPE-seq.[25] We reasoned that the ability of TASQs to isolate G4-DNA deserves both to be investigated and to be compared to that of BG4. We thus report here on these investigations, which provides the very first comparison of the efficiency of the immuno- *versus* chemo-precipitation of G4s, under strictly identical experimental conditions.

Results.

BG4-IP-seq and G4DP-seq protocols. The BG4-IP-seq protocol[15] was implemented with purified rice genomic DNA, after fragmentation (sonication) into 100- to 500-bp sequences in size. These DNA fragments were then thermally treated (10 min at 90 °C, followed by a slow return to 25 °C in G4-stabilzing conditions, 150 mM K^+) in order to maximize G4 folding potentials. BG4-IP-seq thus relies on the use of a FLAG tagged BG4 (expressed from the pSANG10-3F-BG4 plasmid, 3 µg for 5 µg of DNA), followed by the addition of an anti-BG4 Ab (anti-FLAG, 3 µg) and then, protein G-coupled magnetic beads for pulling down the G4/BG4/anti-FLAG/beads complexes. The BG4-IP-seq protocol was then adapted to the TASQ molecular tools: in reference to their use as baits for G4-RNA in the G4RP-seq,[16, 20] we refer to this protocol as G4DP-seq (for G4-DNA precipitation and sequencing). G4DP-seq thus relies on the treatment of the prepared DNA fragments (G4-optimized) with TASQ (either BioTASQ or BioCyTASQ, 100 µM), followed by the addition of streptavidin-coated magnetic beads for pulling down the G4/TASQ/bead complexes. In both instances (BG4-IP-seq and G4DP-seq), thermal elution steps (65 °C twice, 15 min each) followed by DNA purification (PhOH/$CHCl_3$ extraction, then EtOH precipitation) provide samples ready for library preparation.

BG4-IP-seq and G4DP-seq results. The reliability of both BG4-IP-seq and G4DP-seq protocols was assessed by sequencing two independent libraires for each condition. Raw FASTQ reads were aligned to the MSU v7.0 rice genome reference and only reads with a high mapping quality (mapq score >10) were kept for further analyses (**Figure 2A** and Table S1). The use of BG4 led to a high and consistent number of reads (73 to 78,000), with 52,088 common reads; BioTASQ to a variable number of reads (from 51 to 87,000), with 48,722 common reads; BioCyTASQ to a low number of reads (24 to 47,000), with 17,747 common reads. The ratio of common reads (*i.e.*, 71, 94 and 74%) testifies to the good reproducibility of the BG4-IP-seq

and G4DP-seq protocols, which was confirmed by a heat map of Spearman correlation values (Figure S1).

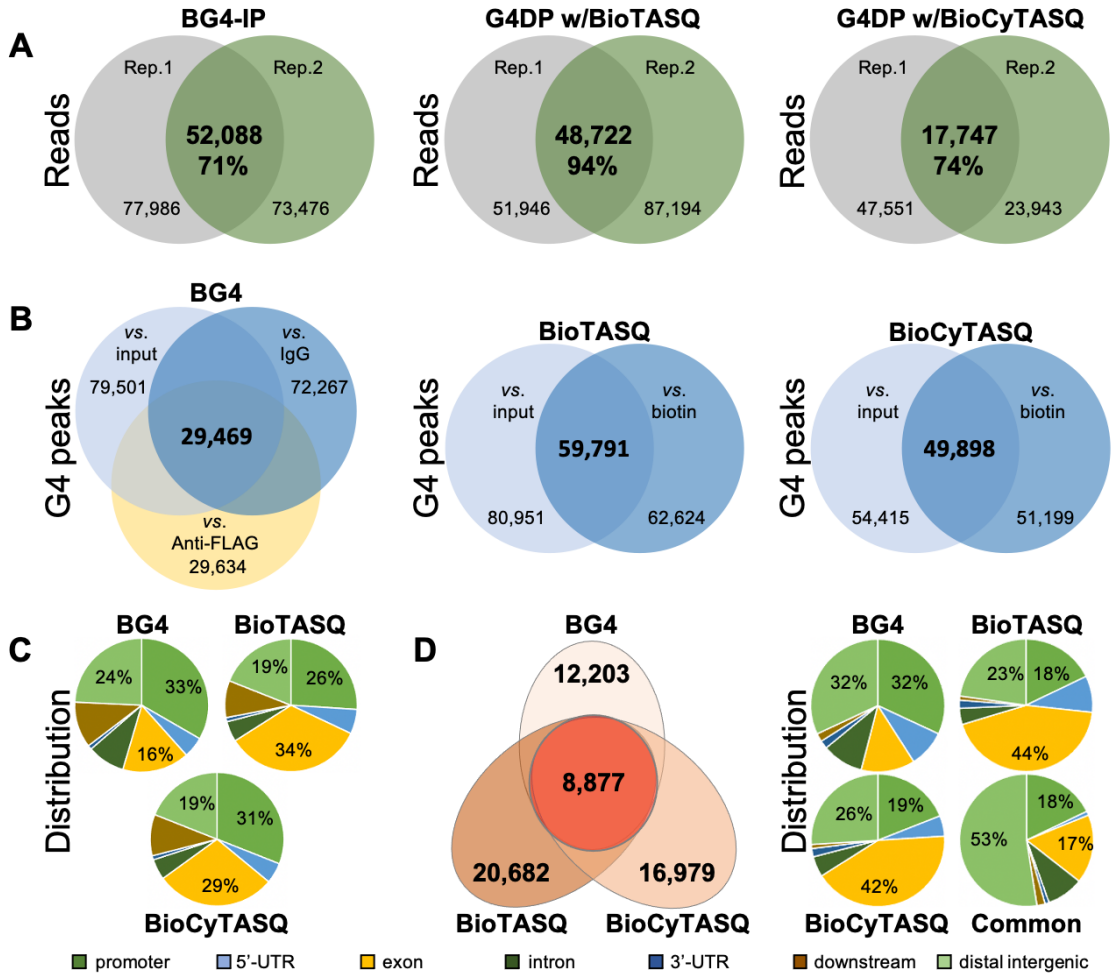

Figure 2. A. Venn plots illustrating the reproducibility of BG4-IP-seq (performed with BG4) and G4DP-seq (performed with either BioTASQ or BioCyTASQ) protocols. **B**. G4 peak calling performed against controls (input, IgG and anti-FLAG for BG4-IP-seq; input and biotin for G4DP-seq) that leads to the identification of high-confidence G4 peaks. **C**. Genomic distribution of the high-confidence G4 peaks. **D**. Pairwise comparison of G4 peaks and genomic distribution of the 12,203 BG4-specific peaks, 20,682 BioTASQ-specific peaks, 16,979 BioCyTASQ-specific peaks and 8,877 common peaks.

Next, clean reads with a length >50 nt were used for G4 peak calling (using MACS2),[26] performed against three controls for BG4 (input, IgG and anti-FLAG alone) and two controls for TASQ (input and biotin). As seen in **Figure 2B**, 29,469, 59,791 and 49,898 high-confidence G4 peaks (*i.e.*, found when compared to all controls) were identified for BG4, BioTASQ and BioCyTASQ, respectively. This indicates that both Ab and ligands capture G4s efficiently, with a somehow better performance of TASQs *versus* BG4 in terms of number of detected peaks. The genomic distributions of these peaks are globally comparable (**Figure 2C**), the most

abundantly detected G4s belonging to promoters, distal intergenic regions and exons, although with some variations in the distribution (33, 24 and 16% for BG4; 26, 19 and 34% for BioTASQ; and 31, 19 and 29% for BioCyTASQ). To better highlight the difference between them, a pairwise comparison (**Figure 2D**) shows that the 8,877 common G4s (detected by both Ab and TASQs, *vide infra*) are primarily distal intergenic G4s (53%), while those enriched specifically by each bait are exonic G4s with TASQs (44 and 42% for BioTASQ and BioCyTASQ, respectively), and equally distributed between promoters and distal intergenic regions with BG4 (32%). These variations might originate in a difference of sequence, that is, of G4 structure, which drives epitope recognition by BG4 and accessibility to the external G-quartet (nature, length and distribution of loops) for TASQs.

BG4-IP-seq and G4DP-seq common G4 peaks. The 8,877 common G4s can be considered as high-confidence G4s that they were trapped and enriched by two orthogonal techniques (antibody- and ligand-based affinity capture), which is quite unique and leaves no doubt about their G4 nature. This was further established by multiple expectation-maximizations for motif elicitation (MEME) analyses,[27] which show that the common G4 motifs correspond mostly to two-quartet G4s of general sequence $G_2N_1G_2N_1G_2N_{1\text{-}2}G_2$ (**Figure 3A**). The pairwise comparison also identifies G4s motifs specifically enriched by BioTASQ (20,682), BioCyTASQ (16,979) and BG4 (12,203) that were similarly MEME-analyzed without revealing any bias for particular G4 structural types (**Figure 3A**). This observation lends credence to the interchangeable use of either capture tools.

To go a step further, we bioinformatically searched for putative G4FS (PG4FS) of general $G_{\geq 2}N_xG_{\geq 2}N_xG_{\geq 2}N_xG_{\geq 2}$ sequence (x between 1 and 12) in the specifically enriched G4 peaks using QuadParser[5] (which searches on both forward and reverse strand) using the published regular expression syntax '([gG]{2,}\w{1,12}){3,}[gG]{2,}'. These analyses reveal a better efficiency of both TASQs (with 62,236 and 83,228 PG4FS identified by BioTASQ and BioCyTASQ, respectively) as compared to BG4 (25,317 PG4FS) (**Figure 3B**), which might originate in the G4-stabilizing properties of TASQs (they do not, however, promote G4 folding,[18] as further discussed below). These elevated numbers originate in both the wide range search ($G_{2+}L_{1\text{-}12}$) and the average length of the G4 peaks (0 - 1,000 bp-long), several PG4FS could thus be found on the same G4 peak. Small molecules can thus capture G4-containing DNA sequences with a high efficiency, even more efficiently than Ab. To further investigate this, the G4 density in

each G4 peak was calculated relative to *ad hoc* controls (random sequences of similar size, n = 100), and this confirms the better performances of TASQs (enrichment score (ES) = 2.63 and 2.68 for BioTASQ and BioCyTASQ, respectively) as compared to BG4 (ES = 1.79) (**Figure 3B**). Of note, a similar analysis performed with the 8,877 common G4 peaks confirms their high G4-content (14,388 PG4FS, ES = 2.12).

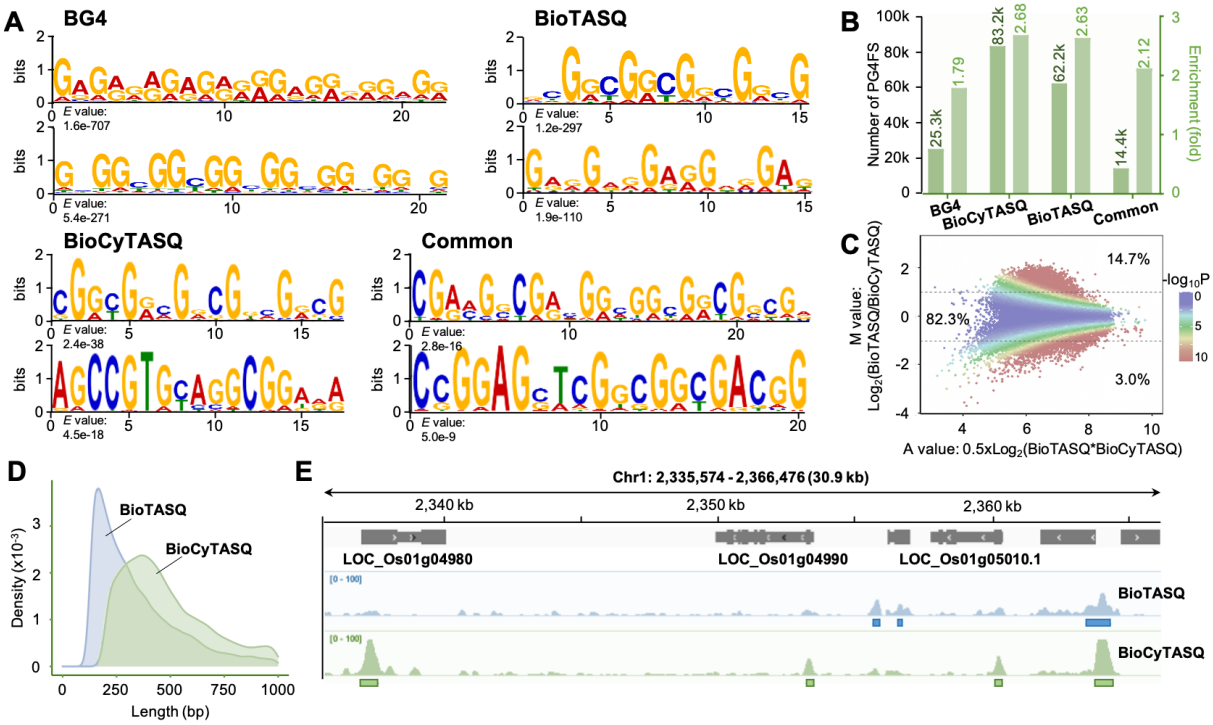

Figure 3. A. Motifs discovery using MEME for BG4-IP-seq and G4DP-seq peak data sets, performed with the 12,203 BG4-specific peaks, 20,682 BioTASQ-specific peaks, 16,979 BioCyTASQ-specific peaks and the 8,877 common peaks, which reveal the presence of repetitive, G-rich motifs. **B**. Number and fold enrichment of putative G4-forming sequences (PG4FSs) identified by QuadParser, which reveal the high likelihood of G4s among the detected BG4-IP-seq/G4DP-seq G4 peaks. **C**. MA-plot illustrating the significant overlap (>80%) between G4 peaks detected using G4DP-seq with both BioTASQ and BioCyTASQ. **D**. Length distribution of G4 peaks detected using G4DP-seq with both BioTASQ (blue) and BioCyTASQ (green). **E.** Example of integrative genomics viewer (IGV) screenshot for a 30-kb window of the rice (*Oryza sativa*) chromosome 1 (Chr1). Tracks are shown for G4DP-seq performed with BioTASQ (blue) and BioCyTASQ (green); G4 peaks are indicated with colored rectangles.

BioTASQ- *versus* BioCyTASQ-based G4DP-seq. It was thus of interest to further compare the properties of the two TASQs. As indicated above, the G4 peaks enriched by both TASQs contain a high number of PG4FS (*ca*. 62 and 83,000), with a very low rate of false positives (<2%, which may contain irregular PG4FSs). The overlap between detected G4 peaks is illustrated by a Bland-Altman plot[28] (or MA-plot: M refers to *minus* and correspond to the log2 ratios on the y axis; A refers to *average* and corresponds to the mean values on the x axis). This visualization compares the agreement between the two G4DP-seq datasets (BioTASQ

versus BioCyTASQ) and reveals a quite homogeneous distribution of the peaks with 82.3% common peaks ($-1 < M < 1$; $-\log_{10}(P) < 2$), along with some BioTASQ- and BioCyTASQ-specific peaks ($5 < A < 8$; $-\log_{10}(P) > 2$; *ca*. 14.7 and 3.0%, respectively, **Figure 3C**).

However, some discrepancies can be noted: indeed, a closer look to the G4 peaks indicates that the distribution of length is wider for BioCyTASQ (average length 550 bp) than for BioTASQ (average length 370 bp), which underlines a significant difference in capture efficiency (**Figure 3D**). Also, a visual inspection of the integrative genomics viewer (IGV) screenshot for a 30-kb region from *Oryza sativa* chromosome 1 shows common and differential G4 peaks captured by BioTASQ (upper rows) and BioCyTASQ (lower rows) but also highlights the better signal-to-noise (S/N) ratio of BioCyTASQ (**Figure 3E**). These differences, which are not easily rationalized but might originate in a difference of molecular flexibility, and thus, of G-quartet stability and accessibility, highlight that BioCyTASQ should be preferred for the capture of G4s *in vitro*.

***in vitro versus in vivo* G4DP-seq**. These results led us to investigate whether TASQs could be suited for the capture of G4s *in vivo*. The G4DP-seq protocol was thus adapted using BioTASQ as bait since it was used for the G4RP-seq, which similarly traps G4-RNA in *in vivo*-like conditions. Rice seedlings were cross-linked using formaldehyde (1%) before being ground in liquid nitrogen to disrupt cell walls and isolate nuclei (several washing steps), which were subsequently lysed for isolating genomic DNA. Cross-linked chromatin was thus fragmented (100- to 500-bp), which was followed by BioTASQ incubation (4 °C, overnight) and addition of streptavidin-coated beads for pulling down the cross-linked G4/TASQ/bead complexes. After thermal elution and cross-link reversal steps, DNA was precipitated for preparation of the sequencing libraries as described above.

The comparison of the *in vitro*/*in vivo* conditions was performed by sequencing two replicated libraries for each condition (Table S2). Clean reads were used for G4 peak calling (MACS2) against two controls (input and biotin): results seen in **Figure 4A** show that very consistent results were obtained in both conditions, with 59,791 and 57,500 high-confidence G4 peaks identified for *in vitro* and *in vivo* G4DP-seq, respectively. Again, a heat map of Spearman correlation values (Figure S2) confirms the high correlation of G4DP-seq replicates, *in vitro* and *in vivo*. The overlap between the G4 peaks is important, with >60% common G4s

(**Figure 4A** and Figure S3) as well as the agreement between the two G4DP-seq data sets, with >70% common peaks (**Figure 4B**).

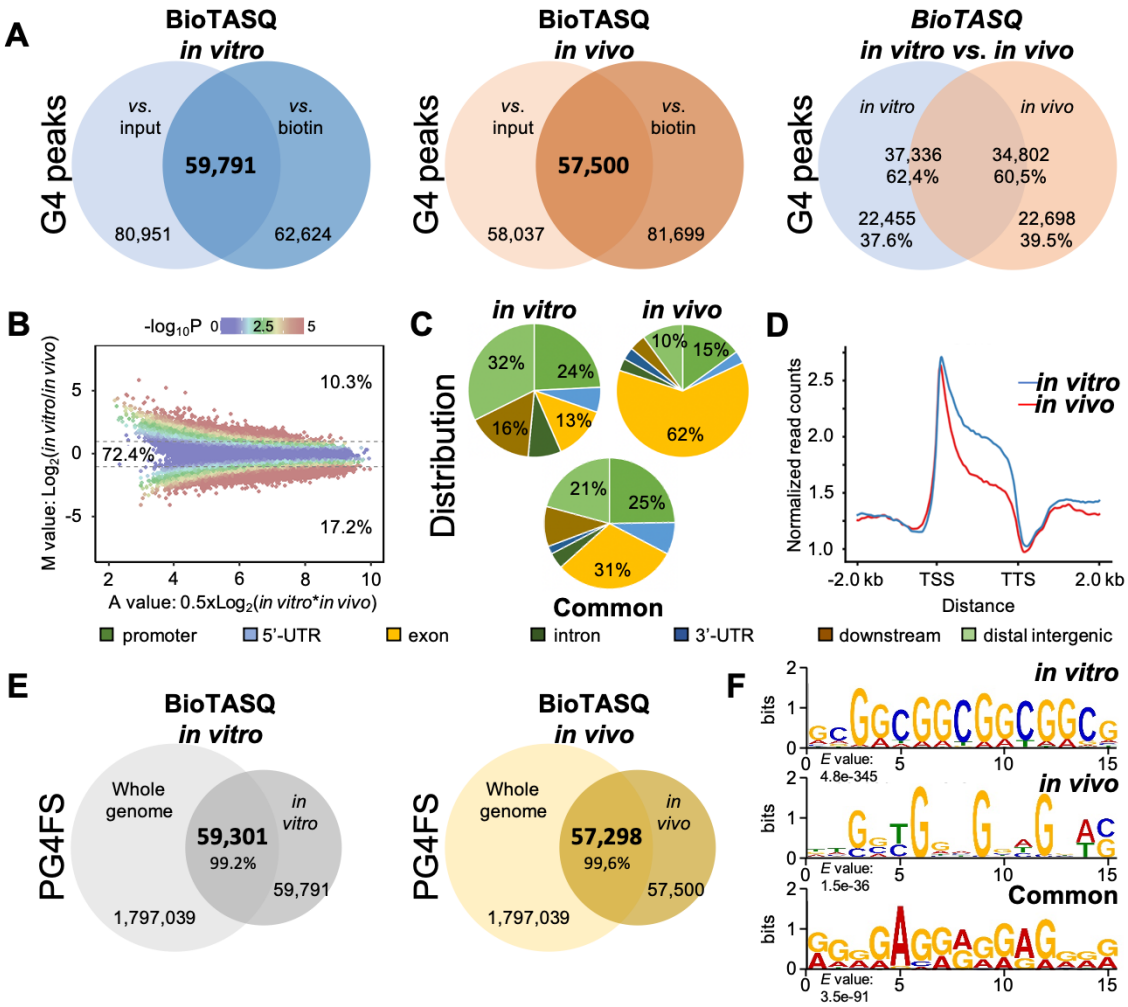

Figure 4. A. Venn plots illustrating the G4 peak calling against two controls (input and biotin) for G4DP-seq performed *in vitro* or *in vivo*, leading to the identification of high-confidence G4 peaks, which are then compared (right plot). **B**. MA-plot illustrating the significant overlap (>70%) between G4 peak data sets of G4DP-seq performed *in vitro* or *in vivo*. **C**. Genomic distribution of the high-confidence G4 peaks detected *in vitro*, *in vivo* or in both conditions (common). **D**. Profiling of *in vivo* and *in vitro* G4DP read counts across ± 2 kb from the transcription start sites (TSSs) to the transcription terminate sites (TTSs) of genes. **E**. Number of PG4FSs identified by QuadParser for G4DP-seq performed *in vitro* or *in vivo,* compared to genome-wide PG4FS. **F**. Motifs discovery using MEME for G4DP-seq peak data set, performed with the 22,455 *in vitro*-specific peaks, 22,698 *in vivo*-specific peaks and the 34,802 common peaks, which reveal the presence of repetitive, G-rich motifs.

However, a significant difference was found when analyzing the genomic distributions of these peaks: while the distribution for *in vitro* G4 peaks was in line with those previously obtained (with notable enrichments in promoters and distal intergenic regions, 24 and 32%, respectively, **Figure 4C**), G4s identified by i*n vivo* G4DP-seq were particularly enriched in exons (62%). This was further demonstrated by profiling the normalized G4 read counts across

genomic regions spanning from 2 kb before the transcription start site (TSS) to 2 kb after the transcription termination site (TTS) of the 55,801 Nipponbare genes, thus encompassing the 7 genomic regions of interest used above (promoters, 5' and 3' UTRs, exons and introns, downstream and distal intergenic regions). The strong enrichment of read counts between TSS and TTS for *in vivo* conditions (**Figure 4D**) is likely due to an over-abundance of exonic G4s. This finding, which strongly suggests key roles of G4s in exons in a functional cellular context, is in line with their recent involvement in alternative splicing events at the transcripts level.[29, 30]

The nature of the identified G4 peaks was further investigated *in silico*: a QuadParser analysis of the high-confidence peaks (59,791 for *in vitro* G4DP-seq; 57,500 for *in vivo* G4DP-seq) revealed a very high G4 motif content, with >99% of these peaks containing the consensus $G_{\geq 2}N_xG_{\geq 2}N_xG_{\geq 2}N_xG_{\geq 2}$ sequence (**Figure 4E**). This high content represents only a minor fraction (*ca*. 3%) of the PG4FS detected genome-wide (*ca*. 1,800,000 putative G4 motifs). The very low difference (<0.1%) between the numbers of G4 peaks identified with BioTASQ *in vivo* and *in vitro* (57,500 and 59,791, respectively) confirms that the TASQ does not promote G4-folding in our conditions. Finally, the G4 motif occurrence within the 22,455 peaks detected *in vitro*, 22,698 peaks detected *in vivo* and the 34,802 common peaks was investigated *via* MEME analyses: as seen in **Figure 4F**, the most enriched motifs were again found to correspond to two-quartet G4s; the particular G-richness observed for the common peaks is interesting as these G4s were trapped and enriched by two orthogonal techniques (*in vitro versus in vivo* G4DP-seq), which is, again, quite unique and leaves no doubt about their G4 nature.

Discussion

Chemical biology and chemical genetics rely on the use of molecular tools that uniquely allow for interrogating biological systems at the cellular level.[31, 32] Scientists have access to a large and diverse portfolio of tools, enabling them to select the most suited molecular devices for their intended applications. Quite often, antibodies (Abs) rank high in this selection, owing to their established high-affinity and -selectivity for their targets. This explains why the antibody BG4 is increasingly used to interrogate G4 biology at the cell/tissue level.[12-15, 33-42] Despite the fact that, by definition, small molecules benefit from more straightforward chemical access, purification and characterization, they still suffer from a globally accepted lower target affinity

and specificity when compared to Abs. However, with a few notable exceptions, including our recent investigations in which we compared chemo- (N-TASQ) *versus* immuno-detection (BG4) of G4s in ALT$^+$ *versus* TERT$^+$ cancer cells,[39] or in *in vitro* single-molecule mechanical unfolding experiments,[43] no direct comparison of the performances of BG4 *versus* multivalent G4 ligands has been attempted in more complicated experimental setups, particularly the chemo-*versus* immuno-precipitation of G4s prior to their identification by sequencing. Our study aims at providing such a comparison.

To this end, the performances of BG4 were compared to that of two small molecules, BioTASQ and BioCyTASQ, under strictly similar experimental conditions. The Ab and TASQs are used to fish G4s out from cell lysates both in *in vitro* conditions where the G4-forming sequences (G4FS) present in the fragmented DNA is properly folded prior to be precipitated by affinity capture, and in *in vivo*-like conditions where naturally folded G4s are fixed (cross-linked) in live cell conditions prior to cell lysis, chromatin fragmentation and affinity capture steps. The Ab/TASQ performances were compared in terms of mapping quality (clean reads), G4 peaks calling along with both the genomic distribution and the G-rich nature (G4 motifs) of the identified G4s. Both Ab- and TASQ-based techniques (BG4-IP-seq and G4DP-seq, respectively) provide an accurate portrayal of G4 biology, notably highlighting a strong bias towards exonic G4s *in vivo*, which indicates the roles that exons might play in a G4-mediated manner in the intricate regulatory networks of plant genetics.

This systematic, multivariate analysis confirms the excellent ability of both molecular devices to isolate G4s for identification purposes. These results show that small molecules (BioTASQ and BioCyTASQ) can thus successfully compete with antibodies (BG4). The performances of TASQs are even better than that of BG4 in terms of number of reads and peaks, which together with their simplest molecular nature makes them ideal tools for chemical biology investigations aiming at deciphering G4 biology in ever greater detail.

Methods

Genetic material. Seeds of rice (*Oryza sativa*) cultivar Nipponbare (Japonica) were pregerminated at 25 °C for 3 d under dark condition. Uniformly germinated seeds were transferred to pots containing the nutrient soil and grown in a greenhouse with an automatically controlled condition: 28 to 30 °C and a 14 h/10 h light/dark cycle. For G4DP-seq *in vivo*: two-week-old rice seedlings were cut into 1-1.5 cm in length and merged into 1 % of

formaldehyde (v/v) in HEPES buffer pH = 8.0 (20 mM HEPES, 1 mM EDTA, 100 mM NaCl and 1 mM PMSF) for cross-linking at 25 °C for 10 min *in vaccuo*. After quenching the excess of formaldehyde by adding 0.125 M final concentration of glycine followed by vacuuming for additional 5 min, the cross-linked leaves were rinsed with autoclaved ddH_2O and air-dried. For G4DP-seq *in vitro* and *in vivo*: cross-linked or native leaves were ground into fine powder using liquid nitrogen; after several washing steps using NIB (nuclear isolation buffer, containing spermine, spermidine and mercaptoethanol), NWB (nuclear washing buffer, containing triton X-100) and NDB (nuclear digestion buffer),[44, 45] chromatin (for cross-linked leaves) and genomic DNA (for native leaves) were isolated and used directly for downstream experiments (or stored at -80 °C for later use).

G4DP-seq protocol *in vitro*. The purified rice genomic DNA was diluted with 1X sonication buffer (50 mM Tris-HCl, 10 mM EDTA and 1 % SDS w/v, pH = 8.0), then fragmented into 100 - 500 bp in size using the water-based Biorupter (Diagnode). A total of 5 µg purified fragmented genomic DNA was diluted in a G4-stabilizing buffer (150 mM KCl and 10 mM Tris-HCl, pH = 7.5), denatured at 95 °C for 10 min, then refolded thanks to a slow return to 25 °C. The refolded DNA was diluted with G4DP buffer (10 mM Tris-HCl, 5 mM EDTA, 150 mM KCl, 0.5 mM DTT, 0.5 % NP-40, pH = 7.4), then incubated with 100 µM of BioTASQ/biotin/BiocyTASQ with a constant rotation at 4 °C overnight. 30 µl of prewashed streptavidin-coupled Dynabeads were added the reaction above for another 4 h-incubation at 4 °C. The washed beads with BiocyTASQ/BioTASQ bound G4 DNA were eluted twice with 200 µl elution buffer (0.1 M $NaHCO_3$ and 1 % SDS w/v) at 65 °C for 15 min each. Precipitated DNA and input DNA was purified using phenol/chloroform extraction followed by cold ethanol precipitation. All libraries were prepared using the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (NEB, E7645S) for paired-end mode sequencing on Illumina NovaSeq platform.

G4DP-seq protocol *in vivo*. The purified cross-linked nuclei were fragmented into sizes ranging from 100 - 500 bp in size in sonication buffer (10 mM Tris-HCl, 5 mM EDTA, 150 mM KCl, 0.5 mM DTT, 0.5 % NP-40 and 0.5 % SDS, pH = 7.4) using the water-based Biorupter. After being centrifuged at 12,000 rpm at 4 °C for 10min, the supernatant containing fragmented chromatin was carefully transferred to a new 1.5 ml tube and kept on ice. The fragmented chromatin was diluted with incubation buffer (10 mM Tris-HCl, 5 mM EDTA, 150 mM KCl, 0.5

mM DTT and 0.5 % NP-40, pH = 7.4) until the final concentration of SDS was below 0.1%. After keeping 1/10 volume of diluted supernatant as input, the remaining volume was incubated with 30 µl of BioTASQ/biotin at 4 °C overnight, then incubated with 30 µl of washed Dynabeads for another 4 h at 4 °C. The washed BioTASQ-bound DNA was eluted with 200 µl elution buffer twice at 65 °C for 15 min each, then reverse cross-linked at 65 °C overnight. The de-cross-linked DNA was purified using phenol/chloroform extraction and cold ethanol precipitation. Precipitated DNA and input DNA was used for library preparation as described above.

Analysis of sequencing data. Raw sequencing data were quality-checked and cleaned using fastp[46] (version 0.21.0). All cleaned reads were aligned to the MSU v7.0 reference genome http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/) using BWA[47] (mem algorithm, version 0.7.17) with default parameters. SAMtools[48] (version 1.5, option -markdup) was used for complete removal of any PCR duplicates. MACS2[26] (version 2.1.1) was used for G4 peak calling using reads with alignment length greater than 50. The command and parameters for G4 peak calling were: macs2 callpeak -g 3.8e+8 -f BAM --nomodel -q 0.01. The input or biotin-precipitated library were used as control during G4 peak calling. Biologically replicated G4 peaks were considered as G4 peaks with high confidence (command intersect of the bedtools package). The plotCorrelation program of deepTools was used for calculating the Spearman's rank correlation coefficients between biological replicates.

Motif prediction. G4 motifs within G4 peaks were predicted using MEME-ChIP (http://meme-suite.org/tools/meme-chip)[27] with the following parameters: minimum width 5 bp, maximum width 25 bp. Only the top significantly enriched motifs (*i.e.*, with the highest *E*-values) are shown in Figures 3 and 4.

PFQSs identification and fold-enrichment analyses. Putative G4-forming sequences (PG4FSs) were identified by screening the whole genome sequences using QuadParser (https://github.com/dariober/bioinformaticscafe/blob/master/fastaRegexFinder.py). The fold-enrichment of PG4FSs ($G_{2+}L_{1-12}$) was calculated relative to random controls across the genome (bedtools shuffle command, observed values divided by average of 100 randomizations values).

Data availability. The BG4-IP-seq data set is available at GSE132775; the G4DP-seq data sets generated in this study are available at the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE200171 (secure token: mxwdquwcljuhvaj)

Acknowledgements

We thank the Bioinformatic Center in Nanjing Agricultural University for providing facilities to assist sequencing data analysis. This research was supported by grants from the National Natural Science Foundation of China (32070561 and U20A2030) and the Fundamental Research Funds for the Central Universities (KYZZ2022001). This work was also supported by the CNRS, the Agence Nationale de la Recherche (ANR-17-CE17-0010-01), and the European Union (PO FEDER-FSE Bourgogne 2014/ 2020 programs, FEDER no. BG0021532). We thank Judy M. Y. Wong (UBC Vancouver, CA) for critical reading of the manuscript.

Authors contribution

Y.F. developed G4DP-seq and performed the experiments; Z.L. designed and implemented the bioinformatic analysis; F.R.S. synthesized both BioTASQ and BioCyTASQ; I.V. designed and optimized TASQs' design and synthesis; W.Z. and D.M. designed the study, interpreted the data and wrote the manuscript.

Additional information

Supplementary information: Summary of *in vitro* BG4-IP-seq and G4DP-seq data information (Table S1) and of *in vitro vs*. *in vivo* G4DP-seq data information (Table S2); heat maps of Spearman correlation values between replicates of BG4-IP-seq and G4DP-seq (Figure S1) and of G4DP-seq *in vitro* and *in vivo* (Figure S2); example of integrative genomics viewer (IGV) screenshot of the rice (Oryza sativa) chromosome 1 (Chr1).

Competing interest

The authors declare no competing interest.

References

1. D. Varshney, J. Spiegel, K. Zyner, D. Tannahill and S. Balasubramanian, *Nat. Rev. Mol. Cell Biol.*, 2020, **21**, 459-474.
2. J. Spiegel, S. Adhikari and S. Balasubramanian, *Trends Chem.*, 2020, **2**, 123-136.
3. S. Burge, G. N. Parkinson, P. Hazel, A. K. Todd and S. Neidle, *Nucleic Acids Res.*, 2006, **34**, 5402-5415.
4. A. K. Todd, M. Johnston and S. Neidle, *Nucleic Acids Res.*, 2005, **33**, 2901-2907.
5. J. L. Huppert and S. Balasubramanian, *Nucleic Acids Res.*, 2005, **33**, 2908-2916.
6. V. S. Chambers, G. Marsico, J. M. Boutell, M. Di Antonio, G. P. Smith and S. Balasubramanian, *Nat. Biotechnol.*, 2015, **33**, 877-881.
7. A. Guedin, J. Gros, P. Alberti and J.-L. Mergny, *Nucleic Acids Res.*, 2010, **38**, 7858-7868.
8. A. Bedrat, L. Lacroix and J.-L. Mergny, *Nucleic Acids Res.*, 2016, **44**, 1746-1759.
9. E. Puig Lombardi and A. Londoño-Vallejo, *Nucleic Acids Res.*, 2019, **48**, 1-15.
10. R. Rodriguez, S. Mueller, J. A. Yeoman, C. Trentesaux, J.-F. Riou and S. Balasubramanian, *J. Am. Chem. Soc.*, 2008, **130**, 15758-15758.
11. G. Marsico, V. S. Chambers, A. B. Sahakyan, P. McCauley, J. M. Boutell, M. D. Antonio and S. Balasubramanian, *Nucleic Acids Res.*, 2019, **47**, 3862-3874.
12. R. Hänsel-Hertsch, D. Beraldi, S. V. Lensing, G. Marsico, K. Zyner, A. Parry, M. Di Antonio, J. Pike, H. Kimura and M. Narita, *Nat. Genet.*, 2016, **48**, 1267-1272.
13. R. Hänsel-Hertsch, J. Spiegel, G. Marsico, D. Tannahill and S. Balasubramanian, *Nat. Protoc.*, 2018, **13**, 551-564.
14. G. Biffi, D. Tannahill, J. McCafferty and S. Balasubramanian, *Nat. Chem.*, 2013, **5**, 182-186.
15. Y. Feng, S. Tao, P. Zhang, F. Rota Sperti, G. Liu, X. Cheng, T. Zhang, H. Yu, X.-e. Wang, C. Chen, D. Monchaud and W. Zhang, *Plant Physiol.*, 2022, **188**, 1632-1648.
16. S. Y. Yang, P. Lejault, S. Chevrier, R. Boidot, A. G. Robertson, J. M. Wong and D. Monchaud, *Nat. Commun.*, 2018, **9**, 4730.
17. I. Renard, M. Grandmougin, A. Roux, S. Y. Yang, P. Lejault, M. Pirrotta, J. M. Y. Wong and D. Monchaud, *Nucleic Acids Res.*, 2019, **47**, 5502-5510.
18. F. Rota Sperti, T. Charbonnier, P. Lejault, J. Zell, C. Bernhard, I. E. Valverde and D. Monchaud, *ACS Chem. Biol.*, 2021, **16**, 905-914.
19. L. Stefan and D. Monchaud, *Nat. Rev. Chem.*, 2019, **3**, 650-668.
20. S. Y. Yang, D. Monchaud and J. M. Y. Wong, *Nat. Protoc.*, 2022, **17**, 870-889.
21. K. Lyu, E. Y.-C. Chow, X. Mou, T.-F. Chan and Chun K. Kwok, *Nucleic Acids Res.*, 2021, **49**, 5426-5450.
22. C. K. Kwok, G. Marsico, A. B. Sahakyan, V. S. Chambers and S. Balasubramanian, *Nat. Meth.*, 2016, **13**, 841-844.
23. J. U. Guo and D. P. Bartel, *Science*, 2016, **353**, aaf5371.
24. X. Weng, J. Gong, Y. Chen, T. Wu, F. Wang, S. Yang, Y. Yuan, G. Luo, K. Chen, L. Hu, H. Ma, P. Wang, Q. C. Zhang, X. Zhou and C. He, *Nat. Chem. Biol.*, 2020, **16**, 489-492.
25. X. Yang, J. Cheema, Y. Zhang, H. Deng, S. Duncan, M. I. Umar, J. Zhao, Q. Liu, X. Cao, C. K. Kwok and Y. Ding, *Genome Biol.*, 2020, **21**, 226.
26. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li and X. S. Liu, *Genome Biol.*, 2008, **9**, R137.
27. P. Machanick and T. L. Bailey, *Bioinformatics*, 2011, **27**, 1696-1697.

28. D. G. Altman and J. M. Bland, *J. R. Stat. Soc., Ser. D Stat.*, 1983, **32**, 307-317.
29. I. Georgakopoulos-Soares, G. E. Parada, H. Y. Wong, R. Medhi, G. Furlan, R. Munita, E. A. Miska, C. K. Kwok and M. Hemberg, *Nat. Commun.*, 2022, **13**, 2404.
30. J. Zhang, S. E. Harvey and C. Cheng, *Nucleic Acids Res.*, 2019, **47**, 3667-3679.
31. K.-H. Altmann, J. Buchner, H. Kessler, F. Diederich, B. Krautler, S. Lippard, R. Liskamp, K. Muller, E. M. Nolan and B. Samorì, *ChemBioChem*, 2009, **10**, 16-29.
32. S. L. Schreiber, *Nat. Chem. Biol.*, 2005, **1**, 64.
33. G. Biffi, M. Di Antonio, D. Tannahill and S. Balasubramanian, *Nat. Chem.*, 2014, **6**, 75-80.
34. G. Biffi, D. Tannahill, J. Miller, W. J. Howat and S. Balasubramanian, *PLoS One*, 2014, **9**.
35. H. Xu, M. Di Antonio, S. McKinney, V. Mathew, B. Ho, N. J. O'Neil, N. D. Santos, J. Silvester, V. Wei, J. Garcia, F. Kabeer, D. Lai, P. Soriano, J. Banáth, D. S. Chiu, D. Yap, D. D. Le, F. B. Ye, A. Zhang, K. Thu, J. Soong, S.-c. Lin, A. H. C. Tsai, T. Osako, T. Algara, D. N. Saunders, J. Wong, J. Xian, M. B. Bally, J. D. Brenton, G. W. Brown, S. P. Shah, D. Cescon, T. W. Mak, C. Caldas, P. C. Stirling, P. Hieter, S. Balasubramanian and S. Aparicio, *Nat. Commun.*, 2017, **8**, 14432.
36. Y. Wang, J. Yang, A. T. Wild, W. H. Wu, R. Shah, C. Danussi, G. J. Riggins, K. Kannan, E. P. Sulman, T. A. Chan and J. T. Huse, *Nat. Commun.*, 2019, **10**, 943.
37. M. Zhang, B. Wang, T. Li, R. Liu, Y. Xiao, X. Geng, G. Li, Q. Liu, C. M. Price, Y. Liu and F. Wang, *Nucleic Acids Res.*, 2019, **47**, 5243-5259.
38. Y.-Z. Xu, P. Jenjaroenpun, T. Wongsurawat, S. D. Byrum, V. Shponka, D. Tannahill, E. A. Chavez, S. S. Hung, C. Steidl, S. Balasubramanian, L. M. Rimsza and S. Kendrick, *NAR Cancer*, 2020, **2**, zcaa029.
39. S. Y. Yang, E. Y. C. Chang, J. Lim, H. H. Kwan, D. Monchaud, S. Yip, Peter C. Stirling and J. M. Y. Wong, *NAR Cancer*, 2021, **3**, zcab031.
40. A. De Magis, M. Kastl, P. Brossart, A. Heine and K. Paeschke, *BMC Biology*, 2021, **19**, 45.
41. T. Masson, C. Landras Guetta, E. Laigre, A. Cucchiarini, P. Duchambon, M.-P. Teulade-Fichou and D. Verga, *Nucleic Acids Res.*, 2021, **49**, 12644-12660.
42. S. Lago, M. Nadai, F. M. Cernilogar, M. Kazerani, H. Domíniguez Moreno, G. Schotta and S. N. Richter, *Nat. Commun.*, 2021, **12**, 3885.
43. P. M. Yangyuoru, M. Di Antonio, C. Ghimire, G. Biffi, S. Balasubramanian and H. Mao, *Angew. Chem. Int. Ed.*, 2015, **54**, 910-913.
44. W. Zhang, Y. Wu, J. C. Schnable, Z. Zeng, M. Freeling, G. E. Crawford and J. Jiang, *Genome Res.*, 2012, **22**, 151-162.
45. W. Zhang and J. Jiang, in *Plant Functional Genomics*, Springer, 2015, pp. 71-89.
46. S. Chen, Y. Zhou, Y. Chen and J. Gu, *Bioinformatics*, 2018, **34**, i884-i890.
47. Y. Jung and D. Han, *Bioinformatics*, 2022, DOI: 10.1093/bioinformatics/btac137.
48. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and G. P. D. P. Subgroup, *Bioinformatics*, 2009, **25**, 2078-2079.

Supplementary Material

Table S1. Summary of *in vitro* BG4-IP-seq and G4DP-seq data information

Sample	Clean reads	Mapping reads	Mapping ratio	Unique reads	Unique reads ratio
Input	43,377,237	43,146,946	99.47%	32,478,543	74.87%
IgGIP	277,616,558	140,225,078	50.51%	26,569,020	9.57%
Anti-FlagIP	238,802,938	124,116,951	51.97%	27,529,338	11.53%
BG4-DNA-IP_rep1	55,500,244	45,680,377	82.31%	21,134,465	38.08%
BG4-DNA-IP_rep2	80,690,945	71,857,860	89.05%	37,099,992	45.98%
Vitro_TASQ_input	42,730,578	42,536,772	99.55%	32,181,480	75.31%
Vitro_TASQ_biotin	78,374,158	74,617,984	95.21%	45,753,270	58.38%
Vitro_BioTASQ_rep1	42,360,919	41,871,151	98.84%	23,027,025	54.36%
Vitro_BioTASQ_rep2	80,821,779	79,940,571	98.91%	42,607,982	52.72%
Vitro_BiocyTASQ_rep1	43,395,247	42,755,522	98.53%	25,040,559	57.70%
Vitro_BiocyTASQ_rep2	23,243,862	22,882,129	98.44%	11,920,947	51.29%

Figure S1. A heat map of Spearman correlation values showing high correlation among biological replicates of BG4-IP-seq and G4DP-seq

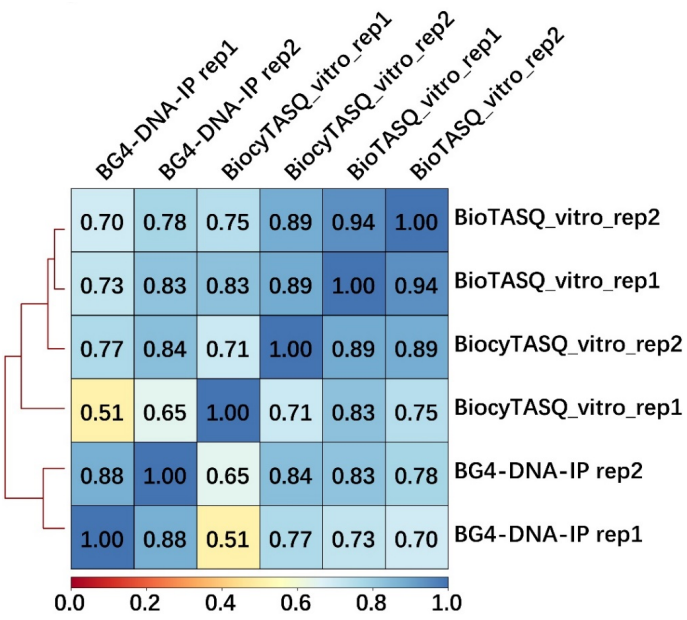

Table S2. Summary of *in vitro vs*. *in vivo* G4DP-seq data information

Sample	Clean reads	Mapping reads	Mapping ratio	Unique reads	Unique reads ratio
BioTASQ_Vivo_input	37,195,245	35,848,553	96.38%	25,674,810	69.03%
BioTASQ_Vivo_biotin	85,963,999	74,278,990	86.41%	49,203,003	57.24%
BioTASQ_Vivo_rep1	51,456,473	50,026,496	97.22%	36,254,883	70.46%
BioTASQ_vivo_rep2	54,370,909	53,855,164	99.05%	39,469,668	72.59%
BioTASQ_Vitro_input	42,730,578	42,536,772	99.55%	32,181,480	75.31%
BioTASQ_Vitro_biotin	78,374,158	74,617,984	95.21%	45,753,270	58.38%
BioTASQ_Vitro_rep1	42,360,919	41,871,151	98.84%	23,027,025	54.36%
BioTASQ_Vitro_rep2	80,821,779	79,940,571	98.91%	42,607,982	52.72%

Figure S2. A heat map of Spearman correlation values showing high correlation among biological replicates of G4DP-seq (*in vitro*/*in vivo*)

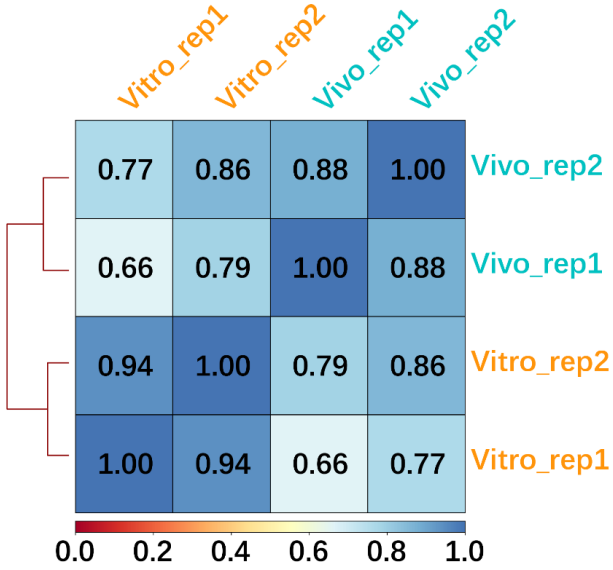

Figure S3. Example of integrative genomics viewer (IGV) screenshot of the rice (*Oryza sativa*) chromosome 1 (Chr1). Tracks are shown for *in vitro* biased peaks, *in vivo* biased peaks and common *in vitro*/*in vivo* peaks by G4DP-seq using BioTASQ.