# SSREnricher: A computational approach for large-scale identification of polymorphic microsatellites based on comparative transcriptome analysis

**Wei Luo** Equal first author, [1] , **Qing Wu** Equal first author, [1] , **Lan Yang** [1] , **Pengyu Chen** [1] , **Siqi Yang** [1] , **Tianzhu Wang** [1] , **Yan Wang** [1] , **Zongjun Du** Corresp. [1]

[1] College of Animal Science and Technology, Sichuan Agricultural University, Chengdu, Sichuan, CHINA

Corresponding Author: Zongjun Du
Email address: 14364@sicau.edu.cn

Microsatellite (SSR) markers are the most popular markers for genetic analyses and molecular selective breeding in plants and animals. However, the currently available methods to develop SSRs are relatively time-consuming and expensive. One of the most factors is low frequency of polymorphic SSRs. In this study, we developed a software, SSREnricher, which composes of six core analysis procedures, including SSR mining, sequence clustering, sequence modification, enrichment containing polymorphic SSR sequences, false-positive removal and results output and multiple sequence alignment. After running of transcriptome sequences on this software, a mass of polymorphic SSRs can be identified. The validation experiments showed almost all markers (> 90%) that were identified by the SSREnricher as putative polymorphic markers were indeed polymorphic. The frequency of polymorphic SSRs identified by SSREnricher was significantly higher ($P < 0.05$) than that of traditional and HTS approaches. The software package is publicly accessible on GitHub [https://github.com/byemaxx/SSREnricher].

# SSREnricher: A computational approach for Large-scale Identification of Polymorphic Microsatellites Based on Comparative Transcriptome Analysis

Wei Luo[#], Qing Wu[#], Lan Yang, Pengyu Chen, Siqi Yang, Tianzhu Wang, Yan Wang, Zongjun Du[*]

College of Animal Science and Technology, Sichuan Agricultural University, Chengdu, 611130, China

[#] These authors contributed equally to this work.

Corresponding Author:
Zongjun Du
Sichuan Agricultural University (Chengdu Campus), 211 Huimin Road, Wenjiang District, Chengdu 611130, PR China.
Email address: 14364@sicau.edu.cn

24


Manuscript to be reviewed

## Abstract

Microsatellite (SSR) markers are the most popular markers for genetic analyses and molecular
selective breeding in plants and animals. However, the currently available methods to develop
SSRs are relatively time-consuming and expensive. One of the most factors is low frequency of
polymorphic SSRs. In this study, we developed a software, SSREnricher, which composes of six
core analysis procedures, including SSR mining, sequence clustering, sequence modification,
enrichment containing polymorphic SSR sequences, false-positive removal and results output
and multiple sequence alignment. After running of transcriptome sequences on this software, a
mass of polymorphic SSRs can be identified. The validation experiments showed almost all
markers (> 90%) that were identified by the SSREnricher as putative polymorphic markers were
indeed polymorphic. The frequency of polymorphic SSRs identified by SSREnricher was
significantly higher ($P < 0.05$) than that of traditional and HTS approaches. The software
package is publicly accessible on GitHub [https://github.com/byemaxx/SSREnricher].

## Introduction

Microsatellites or simple sequence repeats (SSRs) are tandem repeats of one to six nucleotides
present in all eukaryotic genomes. Due to the high mutation rates, co-dominant inheritance, high
abundance, reproducibility of microsatellite-based assays, extensive genome coverage,
chromosome specific location, amenability to automation and high throughput genotyping
(Parida et al., 2009), SSRs have emerged as one of the most popular genetic markers in a wide
range of applications in population genetics (Filippi et al., 2015; Touma et al., 2019),
conservation biology (Han and An, 2009; Kalia et al., 2011; Longwu et al., 2012), marker-
assisted selection (MAS) (Parida et al., 2009; Che et al., 2014; Lv et al., 2014) and other studies
(Moore et al., 1999; Palmer et al., 2013; Parida et al., 2009). In addition, there is increasing
evidence that SSRs can serve a functional role in the regulation of gene expression by affecting
transcription, translational activity, DNA structure, and other metabolic activities (Palmer et al.,
2013).

Despite the broad applicability of SSRs in animal and plant genetics, their development as
genetic markers remains challenging in many species, especially in non-model organisms. The
main reason is that they need to be developed *de novo* in most species being examined for the
first time (Kalia et al., 2011). Traditionally, microsatellite marker development requires the
construction of a genomic library enriched for repeated motifs, isolation and sequencing of
microsatellite-containing clones, primer design, optimization of PCR amplification for each
primer pair, and a polymorphism assessment in a few unrelated individuals (Squirrell et al.,
2003). PCR-based isolation of microsatellite arrays (PIMA) (Lin and Chang, 2013) and fast
isolation by AFLP of sequences containing repeat sites (FIASCO) (Hakki and Akkaya, 2000)
have been devised; however, most of these techniques are inefficient, time-consuming,
multifarious, and expensive (Abdelkrim et al., 2009). A survey of service providers indicated
that high expense (approximately 5,000–10,000 USD) will deliver approximately 10

64  polymorphic loci from anywhere between one to more than three months by traditional cloning
65  approaches (Abdelkrim et al., 2009).
66  With the wide application of high-throughput sequencing (HTS) technologies, especially for
67  transcriptome sequencing, the development of SSR markers using HTS-based approaches has
68  become a feasible alternative for many species (Touma et al., 2019). This method has
69  dramatically reduced the time and cost requirement for large-scale development of SSR markers
70  compared to traditional methods (Wu et al., 2014). Nevertheless, the frequency of polymorphic
71  SSR markers developed by this method has also been very low thus far (Che et al., 2014; Lv et
72  al., 2014; Yang et al., 2014), as most of the loci cannot be effectively applied in genetic analyses
73  and marker-assisted selective breeding. To our best knowledge, there are no reports addressing
74  the low frequency of polymorphic SSR marker development.
75  In this study, we designed a pipeline and integrated it into a user-friendly software called
76  SSREnricher, which can be used to identify polymorphic microsatellites on a large-scale using
77  Trinity (Grabherr et al., 2011) assembled transcript files.
78

79  ## Materials and Methods
80  ### Architectural structure
81  SSREnricher was developed in Python. Polymorphic SSRs enrichment was performed by mining
82  SSR information in transcripts combined with transcript clustering information (Fig. 1). MISA
83  (Thiel et al., 2003) was integrated into the software to mine the SSRs in the transcripts, and CD-
84  HIT (Li and Godzik, 2006) was used to cluster the transcripts. The program composes of six core
85  analysis procedures, including those outlined in Fig. 2.
86  **1) SSR mining:** MISA is integrated into SSREnricher and is used to mine SSR sequence
87  information in transcripts, using the following settings: SSR motifs and the number of repeats
88  shown respectively, monomer-10, dimer-6, trimer-5, tetramer-5, pentamer-5, and hexamer-5.
89  The maximal number of interrupting basepairs in a compound microsatellite is set to 100.
90  **2) Sequence clustering:** CD-HIT-EST is used to cluster transcripts with $\geq 90\%$ similarity to
91  obtain clustering and positive or negative strand information of sequences.
92  **3) Sequence modification:** The negative-strand sequences are converted into reverse
93  complementary sequences.
94  **4) Enrichment containing polymorphic SSR sequences:** The repeat units possessing only
95  one nucleotide are filtered out. Repeat units less than 50bp from the beginning or end of the
96  sequence are also removed.
97  **5) False-positive removal:** The software screens for false-positive based on the enrichment
98  results. The clusters that contain sequences with multiple SSRs loci with the same motif, while
99  others miss some SSRs loci, are removed.
100 **6) Results output and multiple sequence alignment:** Multi-sequence alignment is performed
101 on the obtained clusters to observe the position of the SSRs and generates the result files.

102 The software is supported on macOS and Linux, and is available in two different versions; one
103 with a graphical user interface allowing for easy use, and one in command-line version, allowing
104 for a more flexible and efficient use through the server.

105

106 **Identification of polymorphic SSRs using SSREnricher**
107 To validate our software, we used it to develop SSR markers from rice (*Oryza sativa* L.) and
108 grass carp (*Ctenopharyngodon idella*), the transcriptomes of which are publically available.
109 Analysis in these two species allowed for the validation of our pipeline in both a plant and
110 animal species.
111 The transcriptomes of rice and grass carp were acquired from the sequence read archive (SRA)
112 database of the National Center for Biotechnology Information (NCBI). Three rice
113 (SRR1799209, SRR1974265, and SRR2048540) and two carp (SRR1618540, SRR1618542)
114 transcriptomes obtained from different ecotypes were acquired. The transcriptomes were also
115 assembled using Trinity.

116

117 **Validation experiments**
118 Ten individuals of grass carp and rice were used for SSR validation. For carp, genomic DNA
119 was isolated from tail fin using a salt-extraction method with slight modifications. For rice,
120 genomic DNA was isolated from leaves using the DNeasy Plant mini prep kit (Qiagen). DNA
121 concentration for each sample was determined using NanoDrop (Thermo Scientific). Primers
122 were designed using Primer premier 5 software (Lalitha, 2000). All the PCR amplifications were
123 carried out in a 10 μL volume containing 1 μL 10× buffer (with $Mg^{2+}$), 100 μM dNTPs, 0.5 μL
124 primer pairs, 1 U Taq DNA polymerase and 20 ng genomic DNA. The reaction program was 5
125 min at 95 °C, followed by 30 cycles of 30 s at 94 °C, 30 s at annealing temperature 57 °C, 30 s at
126 72 °C and a final extension at 72 °C for 8 min; then stored at 4 °C. The PCR products were
127 separated by size with Genescan-400 HD size ladders (Applied Biosystems; Foster City, USA),
128 by capillary gel electrophoresis using the ABI 3730 Genetic Analyser (Applied Biosystems). The
129 peak heights and fragment sizes were analyzed using GeneMarker software (Holland and Parson,
130 2011).

131

132 **Results**
133 Three rice and two carp transcriptomes were used to enrich for polymorphic SSRs. The total
134 number of SSRs identified by MISA (non-including mononucleotides) was 9,901 in rice and
135 16,700 in carp (Table 1). After enrichment using our pipeline, a total of 480 and 680 putatively
136 polymorphic SSRs were obtained, respectively. The putative polymorphic SSRs enriched by this
137 program were showed in Appendix file 1. Thirty of rice and 20 loci of grass carp were randomly
138 selected for primer design to test the polymorphism of potential SSRs. After PCR amplification
139 and agarose electrophoresis, 27 (90%) and 20 (100%) of the primers could amplify the sequence
140 with distinct product bands in rice and grass carp, respectively. Of which,  26 (96.30%) and 19

141  (95.00%) loci showed to be polymorphic. The detail information of polymorphic SSRs validated
142  by experiments was showed in Appendix file 2.
143  We compared the HTS approach (Wang et al., 2011) and our pipeline for polymorphic SSRs
144  detection using different transcriptomes in grass carp and rice. The ratio of polymorphic SSRs
145  was raised more than three times. The highest frequency of polymorphic SSRs developed using
146  genomic sequences in rice was 73.6% (Zheng et al., 2008), while our pipeline provided a
147  frequency of up to 95%.
148  We randomly recorded the polymorphic marker frequency in different species, as determined by
149  different methods, including the traditional approach and HTS. The frequency of polymorphic
150  SSRs identified by the traditional clone approach ranged from 5.92% (Estuarine tapertail
151  anchovy (*Coilia nasus*)) to 30.00% (*Pseudogyrincheilus prochilus*) with a mean of 12.04%,
152  while the frequency of polymorphic SSRs divided from transcriptome sequenced by HTS ranged
153  from 9.20% (Pigeonpea (*Cajanus cajan*)) to 73.60% (Rice (*Oryza sativa*)) with a mean of
154  37.93%. The validation experiment showed that the frequency of polymorphic SSRs identified
155  by SSREnricher was above 90% for both species, with a mean of 93.93%. One-way ANOVA
156  indicated that the frequency of polymorphic SSRs identified by this method was significantly
157  higher than that identified by other methods ($P < 0.05$) (Fig. 3).
158

## Discussion

159
160  As the previous reports, the frequency of polymorphic SSR marker development by traditional
161  methods and HTS approaches is relatively low (usually, the frequency ranges from 10% to 70%).
162  Thus, these approaches require a significant amount of time and money for primer design and
163  experimental validation by identifying and separating the many "monomorphic" SSRs from the
164  minority of the polymorphic ones (Kalia et al., 2011; Tang et al., 2008). In the present study, we
165  developed a software, SSREnricher, which can significantly increase the frequency of
166  polymorphic SSR markers. As demonstrated by the validation experiments in rice and grass carp,
167  the frequency of polymorphic SSR was increased to > 90%.
168  SSREnricher was developed based on the idea of enriching homologous sequences containing
169  the same SSR with a different number of repeats and identifying polymorphic SSRs using
170  computational and bioinformatics analyses. Compared with traditional methods, SSREnricher
171  have eliminated the most intensive wet lab steps in a computational way, significantly reducing
172  the running time. In addition, the cost of our method remains much lower than that of
173  conventional methods. Similar to HTS, the most costly step of SSR marker development is the
174  *de novo* transcriptome sequencing when there is no available online sequence data in public
175  databases.
176  The factors that affect the frequency of polymorphic SSRs enriched by the SSREnricher are: 1)
177  genetic differentiation between samples. In theory, the larger genetic differentiation between
178  samples detected by this program, the more putative polymorphic SSRs will be enriched. Thus, it
179  is best not to use individuals with very close relationships, such as full- and half-sib families, to
180  enrich polymorphic SSRs using this program. 2) The specificity of the species itself. The

181  frequency of identifying polymorphic SSRs depends on the species and the genotypes used for
182  the evaluation (Tang et al., 2008). According to a survey of SSR development using expressed
183  sequence tags (ESTs) in more than fifty species, the frequency of polymorphic SSRs varies
184  dramatically (Wu et al., 2014). As previously reported, microsatellite genesis is an evolutionarily
185  dynamic process and has proven to be exceedingly complex, and any change in SSRs resulting in
186  increase or decrease in repeat numbers is associated with mutation rate (Ellegren, 2004; Pearson
187  et al., 2005).
188  With the rapid development of sequencing technologies, the cost and time have been greatly
189  reduced, allowing for widespread use of high-throughput sequencing methods. The number of
190  sequenced and deposited transcriptomes in public databases for many economically important
191  species is continuously increasing. As a consequence, the large-scale development of SSR
192  markers at a low cost and within a short timescale is becoming a feasible task. Thus,
193  development of polymorphic SSR markers using SSREnricher has attractive potential
194  applications in the fields of genetic analyses and molecular selective breeding in plants and
195  animals.
196

## Conclusions

198  Almost all tested SSRs predicted to be polymorphic by the SSREnricher are indeed polymorphic,
199  indicating development of polymorphic SSR markers by the SSREnricher is efficient and
200  reliable. The frequency of polymorphic SSR markers can be significantly increased, and thus the
201  time and cost can be notably decreased by using the SSREnricher.
202

## References

204  Abdelkrim, J., Robertson, B.C., Stanton, J.A.L., Gemmell, N.J., 2009. Fast, cost-effective
205      development of species-specific microsatellite markers by genomic sequencing.
206      Biotechniques. https://doi.org/10.2144/000113084
207  Chabot, C.L., Nigenda, S., 2011. Characterization of 13 microsatellite loci for the tope shark,
208      Galeorhinus galeus, discovered with next-generation sequencing and their utility for eastern
209      Pacific smooth-hound sharks (Mustelus). Conserv. Genet. Resour.
210      https://doi.org/10.1007/s12686-011-9402-y
211  Che, R., Sun, Y., Sun, D., Xu, T., 2014. Characterization of the miiuy croaker (Miichthys miiuy)
212      transcriptome and development of immune-relevant genes and molecular markers. PLoS
213      One. https://doi.org/10.1371/journal.pone.0094046
214  Csencsics, D., Brodbeck, S., Holderegger, R., 2010. Cost-effective, species-specific
215      microsatellite development for the endangered dwarf bulrush (Typha minima) using next-
216      generation sequencing technology. J. Hered. https://doi.org/10.1093/jhered/esq069
217  Dutta, S., Kumawat, G., Singh, B.P., Gupta, D.K., Singh, S., Dogra, V., Gaikwad, K., Sharma,
218      T.R., Raje, R.S., Bandhopadhya, T.K., Datta, S., Singh, M.N., Bashasab, F., Kulwal, P.,
219      Wanjari, K.B., K Varshney, R., Cook, D.R., Singh, N.K., 2011. Development of genic-SSR

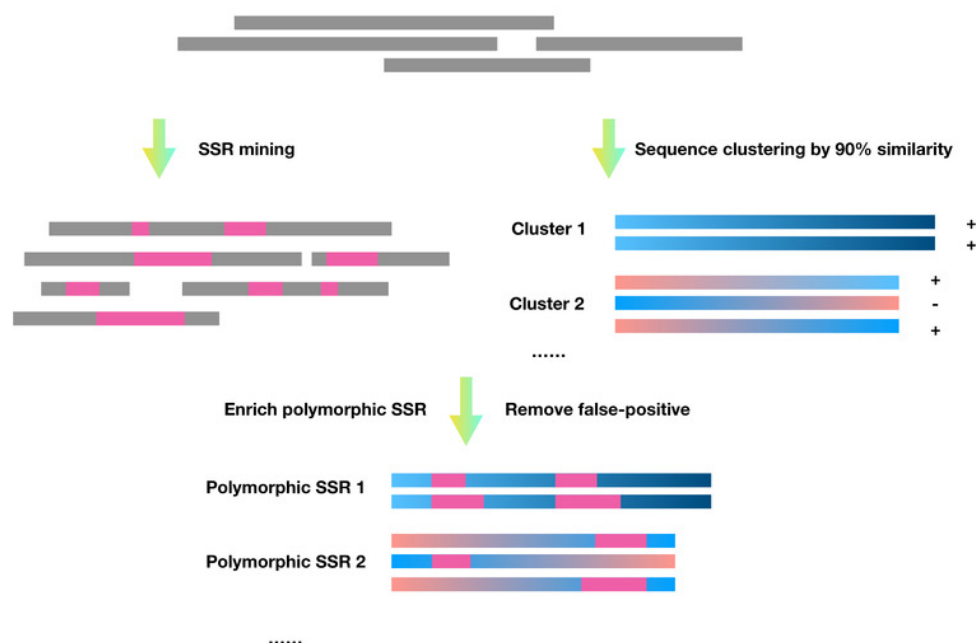
220       markers by deep transcriptome sequencing in pigeonpea [Cajanus cajan (L.) Millspaugh].
221       BMC Plant Biol. https://doi.org/10.1186/1471-2229-11-17
222  Ellegren, H., 2004. Microsatellites: Simple sequences with complex evolution. Nat. Rev. Genet.
223       https://doi.org/10.1038/nrg1348
224  Filippi, C. V., Aguirre, N., Rivas, J.G., Zubrzycki, J., Puebla, A., Cordes, D., Moreno, M. V.,
225       Fusari, C.M., Alvarez, D., Heinz, R.A., Hopp, H.E., Paniego, N.B., Lia, V. V., 2015.
226       Population structure and genetic diversity characterization of a sunflower association
227       mapping population using SSR and SNP markers. BMC Plant Biol.
228       https://doi.org/10.1186/s12870-014-0360-x
229  Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X.,
230       Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A.,
231       Rhind, N., Di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N.,
232       Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a
233       reference genome. Nat. Biotechnol. https://doi.org/10.1038/nbt.1883
234  Hakki, E.E., Akkaya, M.S., 2000. Microsatellite isolation using amplified fragment length
235       polymorphism markers: No cloning, no screening. Mol. Ecol.
236       https://doi.org/10.1046/j.1365-294X.2000.11143.x
237  Han, J., An, L.Z., 2009. Isolation and characterization of microsatellite loci in Peganum harmala
238       (Peganaceae), an important resist-drought and medicinal plant. Conserv. Genet.
239       https://doi.org/10.1007/s10592-009-9849-5
240  Holland, M.M., Parson, W., 2011. GeneMarker® HID: A Reliable Software Tool for the
241       Analysis of Forensic STR Data. J. Forensic Sci. https://doi.org/10.1111/j.1556-
242       4029.2010.01565.x
243  Iorizzo, M., Senalik, D.A., Grzebelus, D., Bowman, M., Cavagnaro, P.F., Matvienko, M.,
244       Ashrafi, H., Van Deynze, A., Simon, P.W., 2011. De novo assembly and characterization of
245       the carrot transcriptome reveals novel genes, new markers, and genetic diversity. BMC
246       Genomics. https://doi.org/10.1186/1471-2164-12-389
247  Kalia, R.K., Rai, M.K., Kalia, S., Singh, R., Dhawan, A.K., 2011. Microsatellite markers: An
248       overview of the recent progress in plants. Euphytica. https://doi.org/10.1007/s10681-010-
249       0286-9
250  Lalitha, S., 2000. Primer Premier 5. Biotech Softw. Internet Rep.
251       https://doi.org/10.1089/152791600459894
252  Li, W., Godzik, A., 2006. Cd-hit: A fast program for clustering and comparing large sets of
253       protein or nucleotide sequences. Bioinformatics.
254       https://doi.org/10.1093/bioinformatics/btl158
255  Lin, H.S., Chang, S. Bin, 2013. PCR-based isolation of microsatellite arrays (PIMA). Methods
256       Mol. Biol. https://doi.org/10.1007/978-1-62703-389-3-3
257  Longwu, G., Cuiyun, L., Guangxiang, T., Chao, L., Wei, X., 2012. Development and
258       characterization of twenty microsatellite markers for the endangered fish Luciobarbus
259       capito. Conserv. Genet. Resour. https://doi.org/10.1007/s12686-012-9660-3

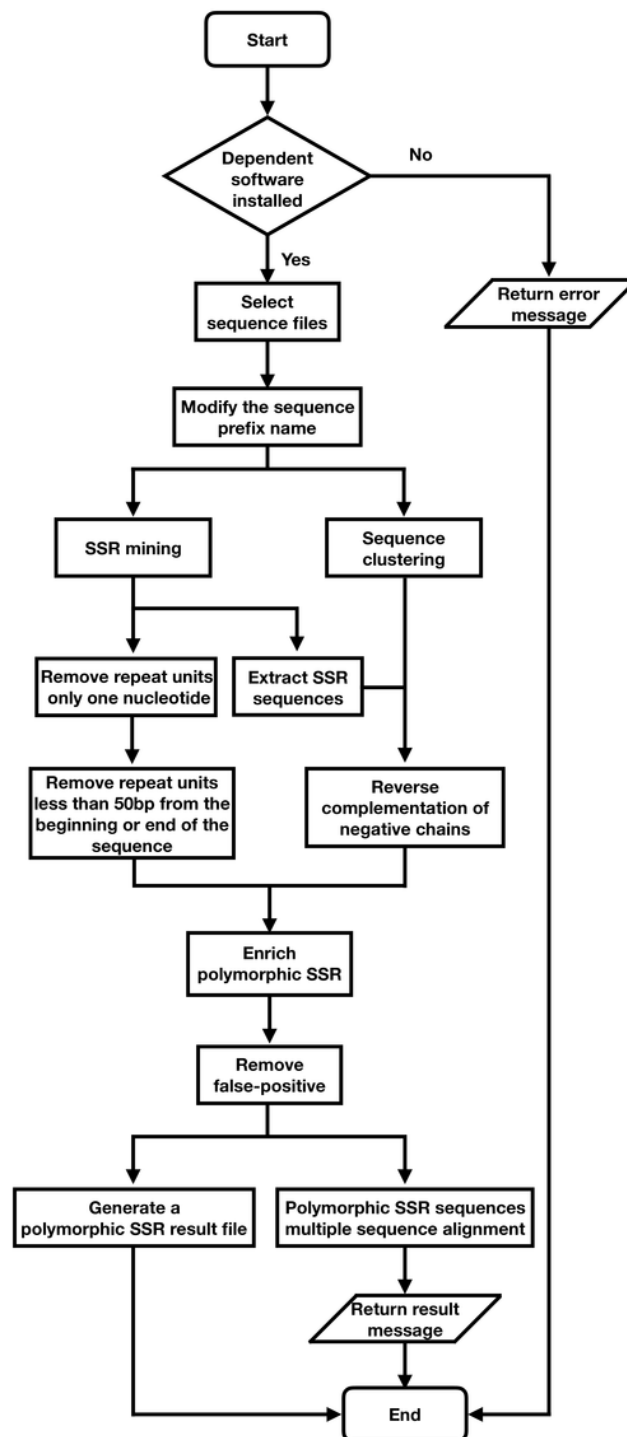
260   Lv, J., Liu, P., Gao, B., Wang, Y., Wang, Z., Chen, P., Li, J., 2014. Transcriptome analysis of the
261         Portunus trituberculatus: De novo assembly, growth-related gene identification and marker
262         discovery. PLoS One. https://doi.org/10.1371/journal.pone.0094055
263   Moore, S.S., Whan, V., Davis, G.P., Byrne, K., Hetzel, D.J.S., Preston, N., 1999. The
264         development and application of genetic markers for the Kuruma prawn Penaeus japonicus,
265         in: Aquaculture. https://doi.org/10.1016/S0044-8486(98)00461-X
266   Palmer, M.E., Lipsitch, M., Moxon, E.R., Bayliss, C.D., 2013. Broad conditions favor the
267         evolution of phase-variable loci. MBio. https://doi.org/10.1128/mBio.00430-12
268   Parida, S.K., Kalia, S.K., Kaul, S., Dalal, V., Hemaprabha, G., Selvi, A., Pandit, A., Singh, A.,
269         Gaikwad, K., Sharma, T.R., Srivastava, P.S., Singh, N.K., Mohapatra, T., 2009. Informative
270         genomic microsatellite markers for efficient genotyping applications in sugarcane. Theor.
271         Appl. Genet. https://doi.org/10.1007/s00122-008-0902-4
272   Pearson, C.E., Edamura, K.N., Cleary, J.D., 2005. Repeat instability: Mechanisms of dynamic
273         mutations. Nat. Rev. Genet. https://doi.org/10.1038/nrg1689
274   Perry, J.C., Rowe, L., 2011. Rapid microsatellite development for water striders by next-
275         generation sequencing. J. Hered. https://doi.org/10.1093/jhered/esq099
276   Santos, B.S., Romana-Eguia, M.R.R., Basiao, Z.U., Ikeda, M., 2015. Development and
277         characterization of nine novel microsatellite markers for the milkfish Chanos chanos.
278         Conserv. Genet. Resour. https://doi.org/10.1007/s12686-014-0393-3
279   Shi, F., Xu, N., Xiong, M., Wang, X., Yang, Z., Que, Y., Zhu, B., Hu, J., Chang, J., 2009.
280         Isolation and characterization of 15 microsatellite loci in an endemic Chinese cyprinid fish,
281         Pseudogyrincheilus prochilus, and their cross-species amplification in two related species.
282         Conserv. Genet. Resour. https://doi.org/10.1007/s12686-009-9092-x
283   Squirrell, J., Hollingsworth, P.M., Woodhead, M., Russell, J., Lowe, A.J., Gibby, M., Powell,
284         W., 2003. How much effort is required to isolate nuclear microsatellites from plants? Mol.
285         Ecol. https://doi.org/10.1046/j.1365-294X.2003.01825.x
286   Tang, J., Baldwin, S.J., Jacobs, J.M.E., van der Linden, C.G., Voorrips, R.E., Leunissen, J.A.M.,
287         van Eck, H., Vosman, B., 2008. Large-scale identification of polymorphic microsatellites
288         using an in silico approach. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-9-374
289   Thiel, T., Michalek, W., Varshney, R.K., Graner, A., 2003. Exploiting EST databases for the
290         development and characterization of gene-derived SSR-markers in barley (Hordeum
291         vulgare L.). Theor. Appl. Genet. https://doi.org/10.1007/s00122-002-1031-0
292   Touma, J., García, K.K., Bravo, S., Leiva, F., Moya, J., Vargas-Chacoff, L., Reyes, A., Vidal, R.,
293         2019. De novo Assembly and Characterization of Patagonian Toothfish Transcriptome and
294         Develop of EST-SSR Markers for Population Genetics. Front. Mar. Sci.
295         https://doi.org/10.3389/fmars.2019.00720
296   Wang, I.J., Summers, K., 2009. Highly polymorphic microsatellite markers for the highly
297         polymorphic strawberry poison-dart frog and some of its congeners. Conserv. Genet.
298         https://doi.org/10.1007/s10592-009-9887-z

299     Wang, J., Yu, L., Bai, J., Fan, J., Quan, Y., 2011. Development of EST-SSR markers and
300         analysis of genetic diversity in five populations of grass carp (Ctenopharyngodon idella).
301         Chinese J. Zool. 46(5), 24–32.
302     Wu, T., Luo, S., Wang, R., Zhong, Y., Xu, X., Lin, Y., He, X., Sun, B., Huang, H., 2014. The
303         first Illumina-based de novo transcriptome sequencing and analysis of pumpkin (Cucurbita
304         moschata Duch.) and SSR marker development. Mol. Breed.
305         https://doi.org/10.1007/s11032-014-0128-x
306     Xie, Z., Xiao, L., Wang, D., Fang, C., Liu, Q., Li, Z., Liu, X., Zhang, Y., Li, S., Lin, H., 2014.
307         Transcriptome analysis of the Trachinotus ovatus: Identification of reproduction, growth
308         and immune-related genes and microsatellite markers. PLoS One.
309         https://doi.org/10.1371/journal.pone.0109419
310     Yang, M., Liang, X.F., Tian, C.X., Gul, Y., Dou, Y.Q., Cao, L., Yu, R., 2012. Isolation and
311         characterization of fifteen novel microsatellite loci in golden mandarin fish (Siniperca
312         scherzeri) Steindachne. Conserv. Genet. Resour. https://doi.org/10.1007/s12686-012-9601-1
313     Yang, Q., Gao, T., Liu, J.X., 2014. Development and characterization of 17 microsatellite loci in
314         an anadromous fish Coilia nasus. Conserv. Genet. Resour. https://doi.org/10.1007/s12686-
315         013-0093-4
316     Zheng, Y., Zhang, G., Lin, F., Wang, Z., Jin, G., Yang, L., Wang, Y., Chen, X., Xu, Z., Zhao, X.,
317         Wang, H., Lu, J., Lu, G., Wu, W., 2008. Development of microsatellite markers and
318         construction of genetic map in rice blast pathogen Magnaporthe grisea. Fungal Genet. Biol.
319         https://doi.org/10.1016/j.fgb.2008.07.012
320

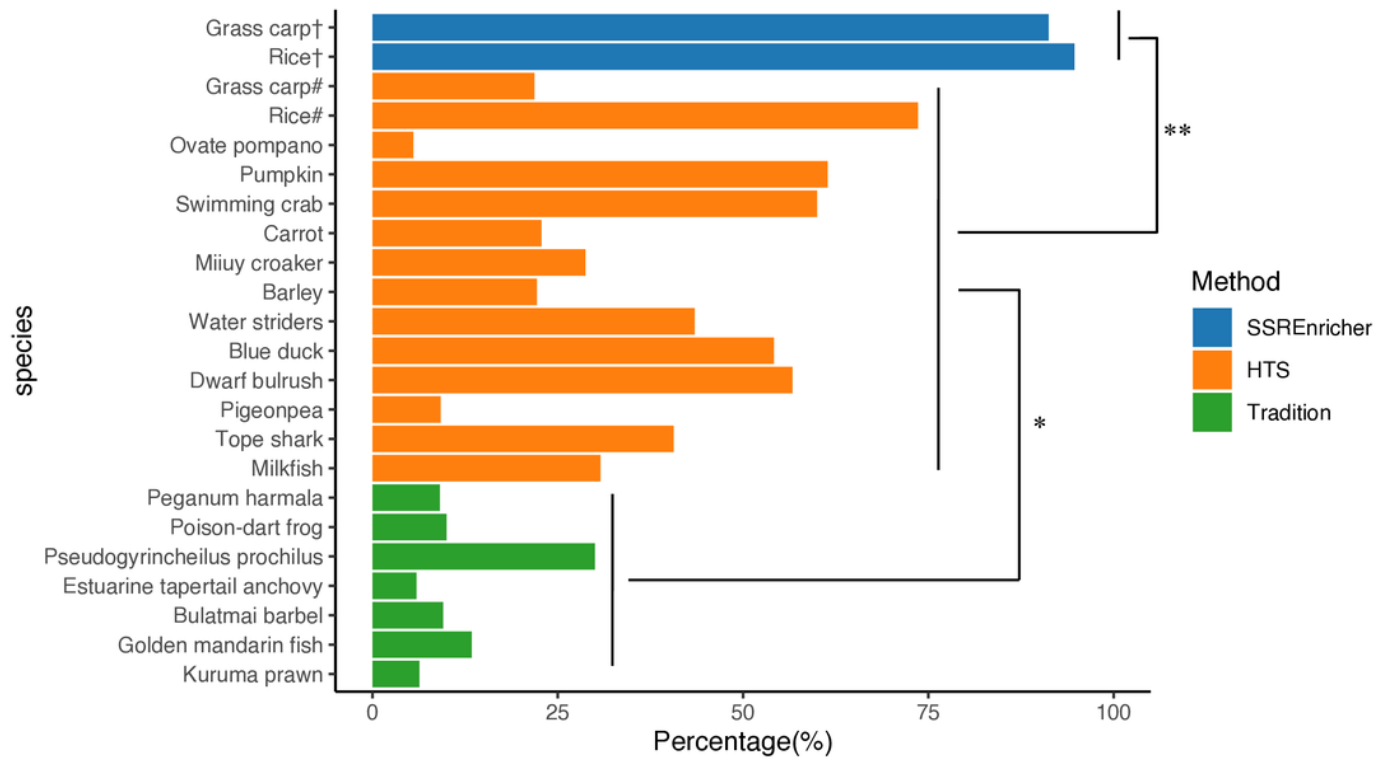
# Figure 1

Schematic diagram of the principles of SSREnricher.

The program first performs SSR mining and sequence clustering on the sequences, and then combines the two results to enrich the polymorphic SSRs.

# Figure 2

Program flow chart of SSREnricher.

# Figure 3

Comparison of polymorphic marker frequency developed by different methods.

# represents SSRs developed by HTS; † represents SSRs developed by this method; * and ** represents significance with $P < 0.05$ and $P < 0.01$ by ANOVA, respectively. The organisms along with their ratio of polymorphic SSR markers used in this figure were as follows: Kuruma prawn (*Penaeus japonicas*)(Moore et al., 1999), Golden mandarin fish (*Siniperca scherzeri*)(Yang et al., 2012), Bulatmai barbel (*Luciobarbus capito*)(Longwu et al., 2012), Estuarine tapertail anchovy (*Coilia nasus*)(Yang et al., 2014), *Pseudogyrincheilus prochilus* (Shi et al., 2009), Poison-dart frog (*Dendrobates pumilio*)(Wang and Summers, 2009), *Peganum harmala* (Han and An, 2009), Milkfish (*Chanos chanos*)(Santos et al., 2015), Dwarf bulrush (*Typha minima*)(Csencsics et al., 2010) , Tope shark (*Galeorhinus galeus*)(Chabot and Nigenda, 2011), Pigeonpea (*Cajanus cajan*)(Dutta et al., 2011), Water striders (*Gerris remigis*)(Perry and Rowe, 2011), Barley (*Hordeum vulgare*)(Thiel et al., 2003), Miiuy croaker (*Miichthys miiuy*)(Che et al., 2014), Carrot (*Daucus carota*)(Iorizzo et al., 2011), Pumpkin (*Cucurbita moschata*)(Wu et al., 2014), Ovate pompano (*Trachinotus ovatus*)(Xie et al., 2014), Swimming crab (*Portunus trituberculatus*)(Lv et al., 2014), Blue duck (*Hymenolaimus malacorhynchos*)(Abdelkrim et al., 2009), Rice (*Oryza sativa*)(Zheng et al., 2008), Grass carp (*Ctenopharyngodon idella*)(Wang et al., 2011).

Number of SSRs, polymorphic SSRs frequency of two verified species.

**Notes:** [1], the mean number of SSRs in the transcriptome (except mononucleotide) SSRs;

[2], the putative polymorphic SSRs enriched by this program were showed in Appendix file 1;

[3], the percentage of primers that can be amplified with clear products bands;

[4], the percentage of SSRs that are polymorphic;

[5], the detail information of polymorphic SSRs validated by experiments was showed in Appendix file 2.

| | Species for verification | |
|---|---|---|
| | Rice | Grass carp |
| Source of the transcriptome | SRR1799209<br>SRR1974265<br>SRR2048540 | SRR1618540<br>SRR1618542 |
| Total number of SSRs in transcriptome[1] (detected by MISA) | 9,901 | 16,700 |
| Number of putatively polymorphic SSRs enriched by this program | 480[2] | 680[2] |
| Number of primers designed for validation experiments | 30 | 20 |
| Number of primers amplified with clear product bands[3] (%) | 27 (90%) | 20 (100%) |
| Number of polymorphic SSRs[4] (%) | 26[5] (96.30%) | 19[5] (95.00%) |