A weighted sequence alignment strategy for gene structure annotation lift over from reference genome to a newly sequenced individual

Baoxing Song^{1,2,3,\$}, Qing Sang^{2,3,\$}, Hai Wang⁴, Huimin Pei³, Fen Wang^{3*}, XiangChao Gan^{2*}

¹ Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA

² Max Planck Institute for Plant Breeding Research, Köln, Germany.

³ The department of life science, Qiannan Normal College for Nationalities, Guizhou
province, China.

⁴ Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing,
China

*To whom correspondence should be addressed.

^{\$}BS and QS contributed equally to this work.

BS: bs647@cornell.edu

QS: qing@mpipz.mpg.de

HW: wanghai01@caas.cn

HP: peihuimin@sgmtu.edu.cn

17 FW: fenmin521@sgmtu.edu.cn

18 XG: gan@mpipz.mpg.de

19

20 **Abstract**

21 Genome sequences and gene structure annotation are very important for genomic analysis,
 22 while only the reference gene structure annotation is widely used for a wide range of
 23 investigations of different natural variation individuals. Herein, we are reporting the software
 24 GEAN which could lift over the reference gene structure annotation to other individuals
 25 belonging to the same or closely related species whose genome sequence was determined by
 26 whole-genome resequencing or *de novo* assembly. We found that inconsistent sequence
 27 alignment makes the coordinate lift over between different individual genomes unreliable,
 28 thus obscuring the lift over of gene structure annotations and genomic variants functional
 29 prediction. We designed a zebraic dynamic programming (ZDP) algorithm by providing
 30 different weights to different genetic features to refine the gene structure lift over. Using the
 31 lift over gene structure annotation as anchors, a base-pair resolution whole-genome-wide
 32 sequence alignment and variant calling pipeline for *de novo* assembly have been
 33 implemented. Taking *Arabidopsis thaliana* as example, we show that the natural variation
 34 alleles expression level of apoptosis death and defence response related genes might could be
 35 better quantified using GEAN. And GEAN could be used to refine the functional annotation
 36 of genetic variants, annotate *de novo* assembly genome sequence, detect syntenic blocks,
 37 improve the quantification of gene expression levels using RNA-seq data and genomic
 38 variants encoding for population genetic analysis. We expect that GEAN will be a standard

39 gene structure annotation lift over and genome sequence alignment tool for the coming age of
40 *de novo* assembly population genetics analysis.

41 **Keywords:** weighted sequence alignment, gene structure annotation, whole genome
42 multiple-sequence alignment, genetic variants uniformization

43 **Introduction**

44 Genome sequence and gene structure annotation are required for modern biological genome
45 research. For large-scale population genetics projects, a high-quality reference genome
46 sequence is usually assembled, and great efforts are exerted to generate the reference gene
47 structure annotation. The genotypic data of a group of taxa are compared with the reference
48 genome sequence. For whole-genome-wide resequencing variant calling projects, single-
49 nucleotide polymorphisms (SNPs), insertions and deletions (INDELs), and other complex
50 variants are generally represented by chromosome, coordinate, reference allele, and
51 alternative allele (Danecek et al., 2011; Gan et al., 2011). While only the gene structure
52 annotation of the reference sequence is generally used for the down-stream analysis e.g. gene
53 expression level using RNA-seq, chip-seq. With the cost reduction of long-read sequencing
54 technology and seeping up of genome assembly (Xiao et al., 2017; Ruan and Li, 2019),
55 increasing natural variation individuals and species is being and will be sequenced with long
56 reads and assembled *de novo* to reveal large-scale structure and rearrangements (Jain et al.,
57 2018; Sun et al., 2018). Gene structure annotations of newly sequenced individuals are very
58 important for many modern analyses, e.g., gene expression quantification, natural variation,
59 population genetics, and strain-specific genes (Zhao et al., 2018; Lilue et al., 2018). While
60 gene structure annotations of *de novo* assemblies have been predicated *ab initio*, and
61 orthologous genes are detected by the alignment of only genetic sequences, which is

independent of the genome sequence alignment (Zapata et al., 2016; Gan et al., 2016; Lilue et al., 2018). Algorithms have been developed to lift over the coordinates between different assembly versions (Hinrichs et al., 2006, 200; Spudich and Fernández-Suárez, 2010; Zhao et al., 2014). Here, we introduced the software GEAN, which could project the reference gene structure annotation to new genome sequence assemblies of the same or closely related species. Using the gene structure annotation as anchors, pipelines to perform base-pair resolution, whole-genome-wide pairwise and multiple sequence alignment (MSA), and variant calling has been implemented.

The lift over of the reference genomic coordinates to the pseudo-genome sequences of re-sequenced individuals could be performed with the variant calling records (Wang et al., 2010; MacArthur et al., 2012) by counting the number of base pairs shifted by the upstream variants. For the *de novo* assembly genomes, GEAN embeds an algorithm, which provides base-pair resolution, long region sequence alignment, and the reference coordinate could be lifted over with the refined sequence alignment. While the same variants might be represented in multiple ways, left alignment and whole-genome-wide multiple sequence alignment have been proposed to solve this problem for population genetics aim (Tan et al., 2015; Song et al., 2018). While the ambiguity of variant representation could cause false-positive open-reading-frame (ORF)-state interruption prediction (Fig. 1). And, there is rare well-designed software to perform genome sequence alignment and lift over the reference gene structure annotation to a new assembly that show considerations for gene structure. Thus, the generation of false-positive ORF-shift predication and improper genotypic variant functional predication is expected.

Here, we provide a semi-global sequence alignment algorithm and software to infer about the gene structure of non-reference accession/line with an algorithm called zebraic dynamic programming (ZDP). Dynamic programming could be accelerated by extending the striped Smith-Waterman (SSW) algorithm (Farrar, 2007), and the accelerated version of ZDP is called zebraic striped dynamic programming (ZSDP). We demonstrate that the ZSDP algorithm result is supported by RNA-seq data. And this method helps to infer the functional impact of variants with the data from *Arabidopsis thaliana* 1001 Genomes project (Alonso-Blanco et al., 2016), gene expression level quantification, annotate *de novo* assembly genome sequence, and perform variant calling for the *de novo* assembly of the genome sequence (Zapata et al., 2016; Gan et al., 2016; Jiao et al., 2017; Springer et al., 2018). Moreover, we are extending this method to other genomic fields.

Fig. 1. Example of inconsistent sequence alignments that affect variants functional inference. (A) Upper panel standard sequence alignment suggests that a 13 bp deletion disturbs the splice site and shifts the ORF of AT5G37190.1. Lower panel alignment, updated using ZSDP, suggests that a 13 bp deletion is located in the intron, and the splice site is conserved. (B) Upper panel shows the gene structure annotation of AT5G37190.1 of a1 allele by coordinate

lift over using standard sequence alignment. Lower panel shows the gene structure annotation updated by the ZSDP algorithm, and the RNA-seq reads mapping support the lower panel.

In the subsequent sections, we demonstrate the performance of GEAN by using *Arabidopsis thaliana*, *Cardamine hirsuta* (Gan et al., 2016), and *Drosophila melanogaster* genome sequences and related evidence. First, by using the *Arabidopsis thaliana* and *Drosophila melanogaster* genomes, we develop concepts that underlie the algorithm of GEAN as a tool to provide gene structure annotation for different inner species individuals. We then consider the inter-species genomes and try to transform the gene structure annotation of *Arabidopsis thaliana* to *Cardamine hirsuta*. We also show that the lift over gene structure annotation could help perform variant calling for the *de novo* assembly genome sequence and whole-genome-wide multiple sequence alignment (MSA).

Lift over of genomic coordinates

Advances in and decreasing costs of high-throughput sequencing have revolutionized the ability to investigate sequence diversity, thereby allowing the international consortium of geneticists to develop foundational resources (Consortium, 2012; Consortium, 2015; Alonso-Blanco et al., 2016, 1; Wang et al., 2018). The compatibility and the ability to integrate different accessions become very interesting. One of the crucial steps is to lift over the reference coordinates, under investigation, to another genome sequence, so that we can compare the orthologous haplotype sequences of interesting regions from two or multiple different individuals.

For the whole-genome resequencing data, the pseudo genome sequence could be obtained by replacing the reference alleles with alternative alleles taking vcf (Danecek et al., 2011) or sdi (Gan et al., 2011) file as input. We embed an algorithm in GEAN to generate the pseudo genome sequence from variant calling result and lift over reference coordinates to the pseudo genome. The lift over of the gene structure could be performed by the lift over of each boundary coordinate.

To lift over the reference coordinates to another *de novo* assembly genome sequence, initial sequence alignment could be performed using the available low-resolution methods for whole genome sequence (Marçais et al., 2018; Li, 2018) or genetic sequence (Li, 2018). Taking low-resolution genome alignment as anchors, GEAN uses a sliding window (Methods) to perform base-pair resolution (Fig. 2), whole-genome wide sequence alignment and lift over the gene structure annotation with the refined sequence alignment.

Fig. 2. Sliding window sequence alignment method implemented in GEAN for the long sequence alignment (Methods).

Mitigation of false deleterious variants caused by inconsistent alignment

A INDEL/SNP ratio of ~25% has been observed in the population genomes of plant, animal and human beings (Rakocevic et al., 2019), and at least 63%-65% INDEL might be affected by inconsistent alignment in a population (Song et al., 2018). As shown in Fig. 1A, the standard sequence alignment based gene structure projection could lead to gene false loss-of-function prediction and improper variants functional prediction. The genetic load of genotypical variants has been analyzed for various purposes by using the data from whole-genome resequencing projects, e.g., GWAS burden test (Song et al., 2018), deleterious mutations (Ramu et al., 2017) and non-synonymous/synonymous mutation ratio for natural selection analysis (Nekrutenko et al., 2002; Liu et al., 2008). Here, we show the number of transcripts, without variants shifting ORF and disturbing splicing sites, could be falsely predicted as loss-of-function.

By using the variant calling result of 1, 211 *Arabidopsis thaliana* accessions (Alonso-Blanco et al., 2016) and 203 *Drosophila melanogaster* lines (Huang et al., 2014; Dembeck et al., 2015) with IMR/DENOM (Gan et al., 2011), we created the pseudo genome sequence of each accession and performed gene structure lift over with simple coordinate lift over firstly. For the lifted protein coding transcripts, which are predicted as loss-of-function, we realigned the structure annotation by using the ZSDP algorithm (Fig. 3). A total of 6,713 transcripts have been realigned as ORF state conserved for *Arabidopsis thaliana* accessions and 1,946 transcripts for *Drosophila melanogaster*. An average of 158 and 47 transcripts have been realigned as ORF state conserved for each *Arabidopsis thaliana* and *Drosophila melanogaster* accession, respectively. We observed negative correlations (p value $< 2.2e-16$) between the number of realigned transcripts and identity by state (IBS) index for both species (Fig. 4),

159 which suggested that more transcripts have been affected by inconsistent alignment between
160 individuals with increased diversity.

161

162

163 Fig. 3. ZSDP sequence alignment methods. When the reference genome sequence is aligned
164 to the target accession genome sequence, different score strategies are used for different
165 reference genetics region to construct the dynamic programming score matrix. The purpose is
166 to align the CDS regions preferentially.

167 To show the advantage the ZSDP approach, we compared the created gene structure
168 annotation file with the RNA-seq data using Cufflinks v2.2.1 (Trapnell et al., 2012). Raw

RNA-seq reads of 728 *Arabidopsis thaliana* accessions released by the *Arabidopsis* 1001
Epigenomes project (Kawakatsu et al., 2016), were and mapped to pseudo genome sequence.
After filtering low-quality accession (less than 3,000,000 RNA-seq reads available, alignment
rate $\leq 90\%$), 697 accessions were left. And the generated bam files were feed into Cufflinks
v2.2.1 to get all the splice site positions. For all the controversial splice sites between simple
coordinate lift over and ZSDP, 41,235 controversial split sites could not be determined by the
RNA-seq reads due to the absence or low expression level, and the RNA-seq data confirmed
23,448 ZSDP alignment splice sites, which is ~ 20 times that of the standard sequence
alignment lift over of splice sites (1,200).

For regions with potential false loss-of-function predication variant records, GEAN realigns
the transcripts, recall the variants, and replace the old variants with the re-aligned variant
records. The newly re-aligned variant records could be used to predict the functional
annotation of genetic variants (Wang et al., 2010).

Fig. 4. Correlation between the number of realigned transcripts by the ZSDP method and IBS index for *Arabidopsis thaliana* (A) and *Drosophila melanogaster* (B).

Non-reference line RNA-seq reads mapping rate could be improved using pseudo genome sequence

For many RNA-seq projects, regardless of what materials is used, the reference genome sequence is usually used for RNA-seq read mapping (Krizek et al., 2016; Kawakatsu et al., 2016). Here, we tested whether mapping read to the pseudo genome sequence could generate an increased mapping rate and if there any gene show different expression level between those two approaches. For this aim we set up a pipeline to generate a compressive gene structure annotation of pseudo genome sequence. Using GFF file from coordinate lift over, the ORF states of the target line could be checked with the inferenced haplotype sequence. Any ORF-state disruption according to the lift over results will be realigned with the ZSDP algorithm.

The ZSDP method could migrate the reference gene structure annotation to the pseudo genome sequence. However, the new accession maybe has some gene structure variation i.e. number of exons that changed, alternative splice site, novel gene. So in GEAN we complemented the gene structure lift over with orthologue-based gene structure annotation (Slater and Birney, 2005) and the results of external annotation method (e.g., *ab initio* annotation (Stanke and Waack, 2003) and transcript assembly (Trapnell et al., 2012)). As illustrated in Fig. S1, the gene structure predicated by the upstream module would be adapted firstly, and the region predicted as ORF-state shift or non-coding would be handled by the

below modules. An average of 346 transcripts was re-annotated by using ZSDP or Exonerate (Slater and Birney, 2005) for each accession.

We mapped RNA-seq reads of 728 *Arabidopsis thaliana* accessions (Kawakatsu et al., 2016) to Col-0 reference genome sequence and pseudo genome sequence separately. The gene structure annotations for each pseudo genome were obtained using the whole gene structure annotation pipeline in GEAN. *Ab initio* gene structure predications were used for the fifth modules of our pipeline. Raw RNA-seq were trimmed and mapped as described above. Mapped reads were counted using HTSeq 0.9.1 (Anders et al., 2015) with default parameters. The mapping rate to the pseudo genome sequence was significantly higher than that to Col-0 reference genome sequence (94.76% vs. 96.33, p value < 2.2e-16, Wilcoxon signed rank test) (Fig. 5).

Fig. 5. Mapping rate of RNA-seq read to pseudo genome (PG) sequence and Col-0 reference genome (RG) sequence.

Due to improved mapping rate, a group of 5,122 genes had significantly higher expression level when quantified by pseudo genome gene structure annotation. And the sequence diversity (measured using pi value) of these genes is significantly higher than that of the whole genome wide background (Wilcoxon test p value < 2.2e-16). GO analysis was conducted using agriGO (Du et al., 2010) with default settings. These results suggested that the advantage of GEAN pipeline for gene expression level quantification is essential for more diverse genes (Fig. 6), especially genes related to cell death, defense response, and immune response (Fig. 7).

Fig. 6. Sequence diversity of genes with significantly different read counts (GDRC) versus the whole-genome wide (WGW) background.

Fig. 7. Directed acyclic graph of GO biological process terms that are enriched (dark orange) in a set of 5,122 genes showing different expression levels in at least 5 out of 697 accessions in which the RNA-seq reads mapping rate was greater than 90 using both Col-0 genome sequence and pseudo genome as reference.

Transformation of the reference gene structure annotation to *de novo* assembly genome sequence

The comparison of the genomic variants of different species and individuals is a very interesting topic. With the reduced cost of long-read sequencing technologies, the genomes of many species have been or will be sequenced. Some inner-species population sequencing projects are moving from short-read to long-read technology. For comparative genomic analysis, aligning the *de novo* assembly to reference accession or well-developed model

organisms is an important step. The genome alignment information could be used to transform the reference gene structure annotation to the newly assembled genome sequence.

We implemented a gene structure annotation function in GEAN by utilizing the genome sequence alignment information. The genome alignment result is usually denoted by a pair of ranges: one is from the reference sequence and the other is from the query genome sequence. ZSDP performs pairwise sequence alignment with a very efficient sliding window method. It lifts the reference gene structure annotation to the query genome sequence and updates the annotation using the ZSDP algorithm.

We tried to transform the gene structure annotation of *Arabidopsis thaliana* Col-0 to *Arabidopsis thaliana* Ler-0 and *Cardamine hirsuta* reference genome sequence. *Arabidopsis thaliana* Ler-0 was assembled into the chromosome level (Zapata et al., 2016). *Cardamine hirsuta* is a species closely related to *Arabidopsis thaliana* phylogenetically, whose genome has been published (Gan et al., 2016). We found that 34,869 out of 39087 protein coding transcripts of *Arabidopsis thaliana* Col-0 accession could be transformed to Ler-0, 32,307 transformed annotations had a conserved ORF state, and 752 of them were re-aligned by the ZSDP algorithm. A total of 14,905 Col-0 protein coding transcripts could be transformed to *Cardamine hirsuta* as conserved ORF-state, while 7,119 of these were re-aligned by the ZSDP algorithm. The higher proportion of transcripts realigned by ZSDP in *Cardamine hirsuta* was confirmed based on the negative correlation between the number of realigned transcripts and IBS index. A dot plot of the transformed gene annotation from Col-0 to *Cardamine hirsuta* was comparable with the Circos plot, showing synteny between the genomes of *Arabidopsis thaliana* and *Cardamine hirsuta* published before (Gan et al., 2016). In addition, a large proportion of genes that did not follow the large synteny fragments were

observed in *Cardamine hirsuta* than in *Ler-0*. Several syntenic analysis modules have been implemented in the GEAN software taking the reference gene structure annotation file and the lift over gene structure annotation file as input (Fig. 8).

Fig. 2. Dot plot of gene position in Col-0 against the position of genes transformed to other *de novo* assemblies. Dots in red color are the result from standard sequence alignment lift over directly, and those in blue color have been realigned by ZSDP algorithm. (A) transform *Arabidopsis thaliana* Col-0 gene structure annotation to *Ler-0* accession genome sequence using standard sequence alignment. (B) transform Col-0 gene structure annotation to *Ler-0* genome sequence of those genes could not be transformed using standard sequence alignment using ZSDP algorithm. (C) investigate syntenic blocks between Col-0 and *Ler-0* using the transformed gene structure annotations. (D) transform *Arabidopsis thaliana* Col-0 gene structure annotation to *Cardamine hirsuta* genome sequence using standard sequence

alignment. (E) transform Col-0 gene structure annotation to *Cardamine hirsuta* genome sequence of those genes could not be transformed using standard sequence alignment using ZSDP algorithm. (F) investigate syntenic blocks between *Arabidopsis thaliana* Col-0 and *Cardamine hirsuta* using the transformed gene structure annotations.

For a comprehensive annotation of *de novo* assembly genome sequence, the GEAN results could be used as a very reliable information source for a complete gene annotation pipeline (Haas et al., 2008).

Base-pair resolution variant calling for *de novo* assembly genome sequence

Variant calling is important for various population genomics analysis. We implemented a pipeline in GEAN to perform whole-genome wide sequence alignment, and variant calling using the lift over gene structure annotation as anchors. In detail, we aligned the gene structure annotation of *de novo* assembly genome sequence with the reference gene structure annotation by using a Needleman-Wunsch algorithm and treating identical gene ID as match. The start and stop codons of matched genes were used as anchors to split the whole genome sequence into fragments, and the sliding window method was used to perform base-pair resolution sequence alignment and variant calling in each fragment. By applying this method on a *Ler-0 de novo* assembly genome sequence (Zapata et al., 2016), 2,664,186 SNPs and 546,778 INDELs were detected, which are higher than those from whole genome resequencing variant calling (693,834 SNPs and 159,350 INDELs) (Song et al., 2018). Except for the probability that *de novo* assembly has a high power of variant calling in highly diverse regions, we classified all the variants into SNP and INDEL only. Hence, for the structure variation, e.g. relocation, could produce a cluster of variant records. Interestingly,

the INDEL/SNP ratios for the two results are comparable with each other. Any two of the reference genome sequence, the variant calling result and *de novo* assembly are enough to inference the third one. As far as we know, GEAN is the first software could call variant in such a comprehensive manner.

Whole-genome wide MSA

For the variant calling of a population of individuals, inconsistent alignments could generate inconsistent variant records. Previously, we showed that this problem could be solved with whole MSA (Song et al., 2018). While, both the left alignment method (Tan et al., 2015) and genome-wide MSA did not consider the gene structure.

GEAN performs MSA for each genetic feature separately. To make is faster, GEAN try to use short and non-overlapping fragments to run the MSA algorithm. Considering that the boundary of genetic features of functional conserved transcripts could be well defined, GEAN performs MSA for short elements (e.g. intron, CDS, intergenic region) directly and only uses an overlapping sliding window to perform MSA only for long elements.

Method

Definitions

The ORF state has defined in the same way as (Song et al., 2018).

Coordination lift over

Considering the presence of INDELs, SVs, and even sequence fragment rearrangements, the coordinates of a certain orthologous haplotype sequence fragments are different from each individual. Lift over is a way of mapping coordinate from one genome assembly to another.

For whole-genome resequencing projects, GEAN performs lift over by counting the number of base pairs that were shifted by the upstream variants. For the *de novo* assembly sequence, considering the similar range entries as input (e.g., Mummer(Marçais et al., 2018), minimap2(Li, 2018, 2)), GEAN generates the base-pair resolution pairwise sequence alignment with a dynamic programming algorithm by using a sliding window, which would be used for the coordinate lift over.

Zebraic dynamic programming

We designed a pairwise sequence alignment algorithm taking gene structure into consideration by extending the available dynamic programming algorithms (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Gotoh, 1982). The two sequences to be compared, query sequence and reference sequence, are defined as $Q=q_1, q_2 \dots q_m$ and $D=d_1, d_2 \dots d_n$, respectively. The length of the query and database sequences are $m=|Q|$ and $n=|D|$, respectively. A scoring matrix $W(q_i, d_j)$ is defined for all residue pairs. The score $W(q_i, d_j) \leq 0$ when $q_i \neq d_j$, and $W(q_i, d_j) > 0$ when $q_i = d_j$. The penalty for starting and extending a gap are defined as G_{init} and G_{ext} , respectively.

As we know the gene structure of the reference sequence, i.e., intron, CDS, start codon, stop codon, and splice sites. So, we use the reference sequence from start codon to stop codon.

And the query sequence is in general extend with variant calling based coordinate lift over, both upstream and downstream by a length of the gene length, to make sure the genetic region include in the selected region. When constructing the score matrix, we initialize the first row and the first column with 0. At the traceback step, we start the tracing from the cell with the largest value in the last column, since we expect the reference sequence could be globally aligned while the query sequence could be locally aligned. And different $W(q_i, d_j)$, G_{init} , and G_{ext} values were used for intron, CDS, splice sites, and start and stop codons separately, so we have all the score strategies as $W(q_i, d_j)_{intron}$, G_{init_intron} , G_{ext_intron} , $W(q_i, d_j)_{CDS}$, G_{init_CDS} , G_{ext_CDS} , $W(q_i, d_j)_{spliceSites}$, $G_{init_spliceSites}$, $G_{ext_spliceSites}$, $W(q_i, d_j)_{start/stopCodon}$, $G_{init_start/stopCodon}$, and $G_{ext_start/stopCodon}$. The protein coding region and splice sites will be highly weighted, and the sequence in those regions will be aligned primarily. Thus, any ambiguity in ORF-states caused by different genetic variance representation could be polished to keep the ORF-states complete.

As we need to determine which gene structure elements of the current j are located and avoid repeating this query process for scoring matrix filling and traceback, we defined a tracking matrix T with the same size as the score matrix, which records the scoring path of each cell of the score matrix. The value of each cell in the track matrix is a certain value SNP, INSERTION, DELETION, or an ambiguous value SNPORINSERTION, SNPORDELETION, INSERTIONORDELETION, SNPORINSERTIONORDELETION.

The alignment scores ending with a gap along D and Q are E calculated using Equation (1) and F calculated using Equation (2), respectively:

$$359 \quad E_{i,j} = \begin{cases} E_{i,j-1} - G_{ext} & \text{if } (T_{i,j-1} \text{ is INSERTION or SNPORINSERTION or} \\ & \text{SNPORINSERTIONORDELETION or} \\ & \text{INSERTIONORDELETION)} \\ H_{i,j-1} - G_{init} & \text{otherwise} \end{cases} \quad (1)$$

$$360 \quad F_{i,j} = \begin{cases} E_{i-1,j} - G_{ext} & \text{if } (T_{i,j-1} \text{ is DELETION or SNPORDELETION or} \\ & \text{SNPORINSERTIONORDELETION or} \\ & \text{INSERTIONORDELETION)} \\ H_{i-1,j} - G_{init} & \text{otherwise} \end{cases} \quad (2)$$

$$M_{i,j} = M_{i-1,j-i} - W(q_i, d_j) \quad (3)$$

361 The alignment score for $H_{i,j}$ where $1 \leq i \leq m$ and $1 \leq j \leq n$ is defined by Equation (4):

$$362 \quad H_{i,j} = \max \begin{cases} E_{i,j} \\ F_{i,j} \\ M_{i,j} \end{cases} \quad (4)$$

363 The matrix T was updated as:

$$364 \quad T_{i,j} = \begin{cases} MATCH, & \text{if } M_{i,j} > E_{i,j} \text{ and } M_{i,j} > F_{i,j} \\ DELETION, & \text{if } F_{i,j} > E_{i,j} \text{ and } F_{i,j} > M_{i,j} \\ INSERTATION, & \text{if } E_{i,j} > F_{i,j} \text{ and } E_{i,j} > M_{i,j} \\ SNPORDELETION, & \text{if } F_{i,j} > E_{i,j} \text{ and } F_{i,j} == M_{i,j} \\ SNPORINSERTION, & \text{if } E_{i,j} > F_{i,j} \text{ and } E_{i,j} == M_{i,j} \\ INSERTIONORDELETION, & \text{if } E_{i,j} == F_{i,j} \text{ and } E_{i,j} > M_{i,j} \\ SNPORINSERTIONORDELETION, & \text{if } E_{i,j} == F_{i,j} \text{ and } E_{i,j} == M_{i,j} \end{cases} \quad (5)$$

365

366 Once $T_{i,j}$ is set as a certain value i.e. SNP, INSERTION, DELETION, SNPORINSERTION,

367 all the continuous ambiguous ancestral cells will be set as certain value, until a certain value

368 being found.

For RAM-saving purposes, instead of keeping the whole $H_{i,j}$ matrix, GEAN only keeps the last row and last column, and the traceback step can be easily performed using the tracking matrix.

ZSDP for gene structure alignment

ZSDP is a speeding up version of ZDP algorithm by intra-sequence parallelization, which parallelize the algorithm by extending the SSW algorithm (Farrar, 2007) to a semi-global sequence alignment manner. The SSW algorithm could be 10 times faster than the standard smith-waterman algorithm using AVX2 and has been widely embedded in several high-throughput sequencing read mapping software, such as Bowtie2 (Langmead and Salzberg, 2012, 2), BWA-SW (Li and Durbin, 2010), and Stampy (Lunter and Goodson, 2011).

In the SSW algorithm, when calculating $H_{i,j}$, the value from the scoring matrix $W(q_i, d_j)$ is added to $H_{i-1,j-1}$. To avoid looking up for $W(q_{i,j})$ for each cell, a query profile parallel to the query is calculated for each possible residue. The query profile is calculated once for each database search. The calculation of $H_{i,j}$ only requires the addition of the pre-calculated score to the previous $H_{i,j}$. The ZSDP method takes a similar approach by pre-calculating a query profile for each gene structure element category separately, and uses a standard method to implement the lazy F evaluation loop.

The SSW method only provides the optimal alignment score but does not report the information necessary to construct the final alignment. The SSW Library reports the detailed alignment by performing SSW twice (get the ending positions by a forward SSW and then generate the beginning position by a backward SSW) and performs a standard smith-

waterman sequences alignment between the beginning and ending position (Zhao et al., 2013), which should be further optimized.

Here, we constructed the $H_{i,j}$ vector by comparing $H_{i-1,j-1}$ and $E_{i,j}$ and stored the $H_{i,j}$ from vHstore into the standard integer scoring matrix. We calculated the $F_{i,j}$ score for each cell of the current column. By comparing the $F_{i,j}$ value with the $H_{i,j}$ value, we updated the $H_{i,j}$ and vHstore with the larger value. The corresponding value in the traceback matrix was also updated. Then, the traceback step of the dynamic programming algorithm, which provides the detailed alignment, could be performed using the traceback matrix.

The computational speeding up is related to the number of cells calculated per CPU register. We used 8, 16, and 32 bits to process the score matrix for sequence with different lengths. So that we could place as many cells as possible into each single instruction multiple data register. Meanwhile, the integer type is wide enough to process the sequence alignment scores. We implemented the striped dynamic algorithm with AVX2. AVX2 is available for most of model Intel processors, whose register is 256 bits wide.

Sequence alignment by using the sliding window

To accelerate the alignment long sequence, we used a sliding window. For each window, we determined the maximum value of the last row or column. The maximum cell would be used as the start cell of the next scoring window. With this strategy, we have a linear computational time costing. Similar with banded alignment, we note that this type of heuristics can fail to give the optimize alignment under some situations. Luckily, the failures are always only present in small local regions. A large sliding window size could be used to

avoid this problem, which is parameter could be passed from command line. This problem could be total avoided, if the sliding window is larger than the sequence being aligned.

Gene expression level quantification

Raw RNA-seq reads were filtered with Trimmomatic (Bolger et al., 2014) (v0.36, LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:6) and mapped to genome sequence using Hisat2 (Kim et al., 2015) (v2.1.0, --max-intronlen 30000).

Ab initio gene structure predications were performed using Augustus (Stanke and Waack, 2003) (--genemodel=complete --maxDNAPieceSize=2000000 --sample=200).

Significantly higher expression qualified using reference genome sequence and using pseudo genome sequence are roughly defined as data out of range lower fence $Q1-5*IQ$ and upper fence $Q3+5*IQ$, $Q1$ is the 25th percentile, $Q3$ is the 75th percentile, and $IQ=Q3-Q1$. And the analysis was performed using R command `boxplot.stats(log(reads_count)ratio, coef=5)$count`.

Software

The gene structure annotation alignment is implemented as the GEAN software. Related functions, such as coordinate lift over any position, obtaining the pseudo genome sequence, and checking the ORF states, are also implemented. Additionally, a function that lifts over all the reference gene structure annotation to other accessions is implemented.

GEAN is written in C++ and parallelized using the C++ Standard Library and freely available at <https://github.com/baoxingsong/GEAN>. A series of examples and tests is embraced in the

distribution. By making the system open source, we hope to encourage others to expand and improve the code base.

Conclusions

As the examples above show, the capabilities of GEAN enable a researcher to transform the gene structure annotation of any two or collections of genomic sequences. This could be used for RNA-seq analysis and perform base-pair resolution, whole-genome-wide sequence alignment by using computer facilities that are widely available today.

Declarations

List of abbreviations

ZDP:(zebraic dynamic programming); SNP:(single-nucleotide polymorphisms); INDEL:(insertions and deletions); MSA:(multiple sequence alignment); ORF:(open-reading-frame); SSW:(striped Smith–Waterman); ZSDP:(zebraic striped dynamic programming); IBS:(identical by state)

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Guizhou Science and Technology Department grants QKH Basic research [2019]1298 and Guizhou Provincial Education Department grants QJH-KY-Z [2018]420 to F.W.

Authors' contributions

BS, QS, FW and XG conceived this software. BS and QS structured the draft and provided final editing. BS coordinated and drafted the manuscript, and synthesized comments provided by all authors. All authors contributed critically important comments. BS implemented the software, HP and FW performed testing and contributed source code. All authors read and approved the final manuscript.

Acknowledgments

We thank Edward S. Buckler and Miltos Tsiantis for the suggestions and comments on the work; Hequan Sun and Bjorn Pieper for discussions; Elad Oren for extending the usage to genome sequence *de novo* assembly quality control. We also acknowledge Sara Miller for the revision.

Authors' information

BS is working as a postdoc at Ed Buckler group at Cornell University. Qing Sang is a postdoc at George Coupland group Max Planck Institute for Plant Breeding Research. FW and HP are working as a lecturer in Qiannan Normal College for Nationalities. XG is a group leader at Max Planck Institute for Plant Breeding Research.

Reference

- Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., et al. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell* 0.
- Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–169.

471 Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina
472 Sequence Data. *Bioinformatics* Advance Access published April 1, 2014,
473 doi:10.1093/bioinformatics/btu170.

474 Consortium, T. 1000 G. P. (2012). An integrated map of genetic variation from 1,092 human
475 genomes. *Nature* 491:56–65.

476 Consortium, T. U. (2015). The UK10K project identifies rare variants in health and disease.
477 *Nature* 526:82–90.

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

481 Dembeck, L. M., Böröczky, K., Huang, W., Schal, C., Anholt, R. R. H., and Mackay, T. F. C.
482 (2015). Genetic architecture of natural variation in cuticular hydrocarbon
483 composition in *Drosophila melanogaster*. *eLife* 4:e09861.

484 Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the
485 agricultural community. *Nucleic Acids Res* 38:W64–W70.

486 Farrar, M. (2007). Striped Smith–Waterman speeds database searches six times over other
487 SIMD implementations. *Bioinformatics* 23:156–161.

488 Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R.,
489 Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., et al. (2011). Multiple reference
490 genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477:419–423.

491 Gan, X., Hay, A., Kwantes, M., Haberer, G., Hallab, A., Ioio, R. D., Hofhuis, H., Pieper, B.,
492 Cartolano, M., Neumann, U., et al. (2016). The *Cardamine hirsuta* genome offers
493 insight into the evolution of morphological diversity. *Nature Plants* 2:16167.

494 Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of*
495 *Molecular Biology* 162:705–708.

496 Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R.,
497 and Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using
498 EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*
499 9:R7.

500 Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans,
501 M., Furey, T. S., Harte, R. A., Hsu, F., et al. (2006). The UCSC Genome Browser
502 Database: update 2006. *Nucleic Acids Res* 34:D590–D598.

503 Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A. M., Turlapati, L.,
504 Zichner, T., Zhu, D., Lyman, R. F., et al. (2014). Natural variation in genome

505 architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines.
506 *Genome Res.* 24:1193–1208.

507 Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D.,
508 Dilthey, A. T., Fiddes, I. T., et al. (2018). Nanopore sequencing and assembly of a
509 human genome with ultra-long reads. *Nature Biotechnology* 36:338–345.

510 Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei,
511 X., Chin, C.-S., et al. (2017). Improved maize reference genome with single-molecule
512 technologies. *Nature* 546:524–527.

513 Kawakatsu, T., Huang, S. C., Jupe, F., Sasaki, E., Schmitz, R. J., Urich, M. A., Castanon, R., Nery,
514 J. R., Barragan, C., He, Y., et al. (2016). Epigenomic Diversity in a Global Collection of
515 *Arabidopsis thaliana* Accessions. *Cell* 166:492–505.

516 Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low
517 memory requirements. *Nature Methods* 12:357–360.

518 Krizek, B. A., Bequette, C. J., Xu, K., Blakley, I. C., Fu, Z. Q., Stratmann, J. W., and Loraine, A. E.
519 (2016). RNA-Seq Links the Transcription Factors AINTEGUMENTA and
520 AINTEGUMENTA-LIKE6 to Cell Wall Remodeling and Plant Defense Pathways. *Plant*
521 *Physiology* 171:2069–2084.

522 Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat*
523 *Meth* 9:357–359.

524 Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
525 34:3094–3100.

526 Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler
527 transform. *Bioinformatics* 26:589–595.

528 Lilue, J., Doran, A. G., Fiddes, I. T., Abrudan, M., Armstrong, J., Bennett, R., Chow, W., Collins,
529 J., Collins, S., Czechanski, A., et al. (2018). Sixteen diverse laboratory mouse
530 reference genomes define strain-specific haplotypes and novel functional loci.
531 *Nature Genetics* Advance Access published October 1, 2018, doi:10.1038/s41588-
532 018-0223-8.

533 Liu, J., Zhang, Y., Lei, X., and Zhang, Z. (2008). Natural selection of protein structural and
534 functional properties: a single nucleotide polymorphism perspective. *Genome Biol*
535 9:R69.

536 Lunter, G., and Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast
537 mapping of Illumina sequence reads. *Genome Res.* 21:936–939.

- 538 MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins,
539 L., Habegger, L., Pickrell, J. K., Montgomery, S. B., et al. (2012). A Systematic Survey
540 of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* 335:823–828.
- 541 Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018).
542 MUMmer4: A fast and versatile genome alignment system. *PLOS Computational*
543 *Biology* 14:e1005944.
- 544 Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for
545 similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*
546 48:443–453.
- 547 Nekrutenko, A., Makova, K. D., and Li, W.-H. (2002). The KA/KS Ratio Test for Assessing the
548 Protein-Coding Potential of Genomic Regions: An Empirical and Simulation Study.
549 *Genome Res* 12:198–202.
- 550 Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., Arsenijevic,
551 V., Nadj, J., Ghose, K., Suci, M. C., et al. (2019). Fast and accurate genomic analyses
552 using genome graphs. *Nature Genetics* 51:354.
- 553 Ramu, P., Esuma, W., Kawuki, R., Rabbi, I. Y., Egesi, C., Bredeson, J. V., Bart, R. S., Verma, J.,
554 Buckler, E. S., and Lu, F. (2017). Cassava haplotype map highlights fixation of
555 deleterious mutations during clonal propagation. *Nat Genet* 49:959–963.
- 556 Ruan, J., and Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. *bioRxiv*
557 Advance Access published January 26, 2019, doi:10.1101/530972.
- 558 Slater, G. S. C., and Birney, E. (2005). Automated generation of heuristics for biological
559 sequence comparison. *BMC Bioinformatics* 6:31.
- 560 Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences.
561 *Journal of Molecular Biology* 147:195–197.
- 562 Song, B., Mott, R., and Gan, X. (2018). Recovery of novel association loci in Arabidopsis
563 thaliana and Drosophila melanogaster through leveraging INDELs association and
564 integrated burden test. *PLOS Genetics* 14:e1007699.
- 565 Springer, N. M., Anderson, S. N., Andorf, C. M., Ahern, K. R., Bai, F., Barad, O., Barbazuk, W.
566 B., Bass, H. W., Baruch, K., Ben-Zvi, G., et al. (2018). The maize W22 genome
567 provides a foundation for functional genomics and transposon biology. *Nature*
568 *Genetics* 50:1282.
- 569 Spudich, G. M., and Fernández-Suárez, X. M. (2010). Touring Ensembl: A practical guide to
570 genome browsing. *BMC Genomics* 11:295.
- 571 Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new
572 intron submodel. *Bioinformatics* 19 Suppl 2:ii215-225.

- 573 Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, H., Zhao, H., Song, W., Zhang, M., Cui, Y., Dong, X., et
574 al. (2018). Extensive intraspecific gene order and gene structural variations between
575 Mo17 and other maize genomes. *Nature Genetics* 50:1289.
- 576 Tan, A., Abecasis, G. R., and Kang, H. M. (2015). Unified representation of genetic variants.
577 *Bioinformatics* 31:2202–2204.
- 578 Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L.,
579 Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis
580 of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* 7:562–578.
- 581 Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic
582 variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164–e164.
- 583 Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.
584 R., Zhang, F., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian
585 cultivated rice. *Nature* 557:43–49.
- 586 Xiao, C.-L., Chen, Y., Xie, S.-Q., Chen, K.-N., Wang, Y., Han, Y., Luo, F., and Xie, Z. (2017).
587 MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule
588 sequencing reads. *Nature Methods* 14:1072–1074.
- 589 Zapata, L., Ding, J., Willing, E.-M., Hartwig, B., Bezdan, D., Jiao, W.-B., Patel, V., James, G. V.,
590 Koornneef, M., Ossowski, S., et al. (2016). Chromosome-level assembly of
591 *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion
592 polymorphisms. *PNAS* 113:E4052–E4060.
- 593 Zhao, M., Lee, W.-P., Garrison, E. P., and Marth, G. T. (2013). SSW Library: An SIMD Smith-
594 Waterman C/C++ Library for Use in Genomic Applications. *PLOS ONE* 8:e82138.
- 595 Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., and Wang, L. (2014). CrossMap: a
596 versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*
597 30:1006–1007.
- 598 Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., et al.
599 (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated
600 and wild rice. *Nature Genetics* Advance Access published January 15, 2018,
601 doi:10.1038/s41588-018-0041-z.

602

603 **Supplementary information:**

604

605

606 Fig. S1. Schematic showing the pipeline for the protein coding gene annotation of the non-
607 reference genome. The gene structure annotation of upstream step would be supplemented
608 with the downstream method.