m^7GDisAI: N7-methylguanosine (m^7G) sites and diseases associations inference based on heterogeneous network

Jiani Ma^1,2, Lin Zhang^1,2*, Jin Chen^1,2, Bowen Song^3, Chenxuan Zang^3 and Hui Liu^1,2

Abstract

Background: Recent studies have confirmed that N7-methylguanosine (m^7G) modification plays an important role in regulating various biological processes and has associations with multiple diseases. Wet-lab experiments are cost and time ineffective for the identification of disease-associated m^7G sites. To date, tens of thousands of m^7G sites have been identified by high-throughput sequencing approaches and the information is publicly available in bioinformatics databases, which can be leveraged to predict potential disease-associated m^7G sites using a computational perspective. Thus, computational methods for m^7G-disease association prediction are urgently needed, but none are currently available at present.

Results: To fill this gap, we collected association information between m^7G sites and diseases, genomic information of m^7G sites, and phenotypic information of diseases from different databases to build an m^7G-disease association dataset. To infer potential disease-associated m^7G sites, we then proposed a heterogeneous network-based model, m^7G Sites and Diseases Associations Inference (m^7GDisAI) model. m^7GDisAI predicts the potential disease-associated m^7G sites by applying a matrix decomposition method on heterogeneous networks which integrate comprehensive similarity information of m^7G sites and diseases. To evaluate the prediction performance, 10 runs of tenfold cross validation were first conducted, and m^7GDisAI got the highest AUC of 0.740(±0.0024). Then global and local leave-one-out cross validation (LOOCV) experiments were implemented to evaluate the model’s accuracy in global and local situations respectively. AUC of 0.769 was achieved in global LOOCV, while 0.635 in local LOOCV. A case study was finally conducted to identify the most promising ovarian cancer-related m^7G sites for further functional analysis. Gene Ontology (GO) enrichment analysis was performed to explore the complex associations between host gene of m^7G sites and GO terms. The results showed that m^7GDisAI identified disease-associated m^7G sites and their host genes are consistently related to the pathogenesis of ovarian cancer, which may provide some clues for pathogenesis of diseases.

Conclusion: The m^7GDisAI web server can be accessed at http://180.208.58.66/m7GDisAI/, which provides a user-friendly interface to query disease associated m^7G. The list of top 20 m^7G sites predicted to be associated with 177 diseases can be achieved. Furthermore, detailed information about specific m^7G sites and diseases are also shown.
Introduction

Over 150 types of RNA modifications have been identified in RNA molecules [1, 2], and N7-methylguanosine (m7G), which refers to methylation of guanosine (G) on position N7, is a typical positively charged modification present in tRNA [3], rRNA [4], mRNA 5′cap [5] and internal mRNA regions [6], playing a critical role in regulating RNA processing, metabolism, and function. As a positively charged RNA modification, m7G could tune RNA secondary structures or protein-RNA interactions through a combination of electrostatic and steric effects [7]. m7G sites in several tRNAs variable loops, which are installed by the heterodimers METTL1-WDR4 in mammals [3], have been reported to stabilize tRNA tertiary fold [8, 9]. m7G sites that install at 5′cap stabilize transcripts against exonucleolytic degradation [10], and modulate nearly every stage of the mRNA life cycle, including transcription elongation [11], pre-mRNA splicing [12], polyadenylation [13], nuclear export [14], and translation [15].

Mutations in m7G methyltransferase are associated with various diseases. To be more specific, a mutation in the methyltransferase complex WDR4 (WD Repeat Domain 4) in humans has been reported to cause primordial dwarfism characterized by facial dysmorphism, brain malformation, and severe encephalopathy with seizures [16, 17]. Lin et al. [18] reported that knockout of the m7G46 tRNA WDR4 in embryonic stem cells impairs neural lineage differentiation and affects translation on a global scale. Besides, overexpression of WDR4 has been discovered to influence learning and memory in Down syndrome [19]. Moreover, the m7G tRNA methyltransferase METTL1 (Methyltransferase like 1) was reported to influence cancer cell viability [20]. Therefore, identification of disease-associated m7G sites will accelerate the understanding of disease pathogenesis at the molecular level, and will further benefit the prognosis, diagnosis, evaluation, treatment, and prevention of human complex diseases. However, it is time-consuming and expensive to explore the association between m7G sites and various diseases by only conducting wet experiments. Fortunately, m7G-MeRIP-Seq [21], m7G-miCLIP-seq [6], and m7G-Seq [21] have generated vast amounts of biological data about m7G, so computational methods are urgently needed to uncover potential disease-associated m7G sites effectively. Researchers can then select the most probable m7G sites and the host genes of these sites for further analysis, streamlining their wet-lab experiments. To our knowledge, no computational models for finding disease-associated m7G sites have been developed.

In this study, we extracted 768 validated associations among 741 m7G sites and 177 diseases from m7GHub to construct the m7G disease association dataset [22]. Then we proposed a heterogeneous network-based m7G-disease associations inference method m7GDisAI to prioritize candidate m7G sites for a disease of interest. Furthermore, experiments of cross validation and case study on ovarian cancer have been carried out to prove the effectiveness and stability of our method. To facilitate the exploration and direct query of our predicted results, we developed an online database m7GDisAI. The website hosts the top 20 m7G sites predicted to be associated with 177 diseases with high prediction scores and supports queries with diseases which you are interested. The m7GDisAI website is freely available at http://180.208.58.66/m7GDisAI/.
Implementation

Datasets

Source of datasets

m7GHub is a comprehensive m7G online platform, which deciphers the location, regulation, and pathogenesis of m7G modification [22]. It consists of four parts, including m7GDB, m7GFinder, m7GSNPer, and m7GdiseaseDB. It provides 69,159 m7G sites which are classified into three confidence levels: high confidence level sites reported by m7G-seq, medium confidence level sites reported by m7G-MeRIP-Seq as well as m7G-miCLIP-Seq, and low confidence level sites predicted by m7GFinder. As a sub-part of m7GHub, m7GDiseaseDB collects 1218 disease-associated genetic variants that may lead to gain/loss of m7G sites, with implications for disease pathogenesis involving m7G RNA methylation. It provides us sufficient information to construct the m7G-variant dataset and further build the m7G-disease association dataset.

m7G-variant dataset

In the m7G-variant dataset, m7G-associated variants refer to those mutated at or close to G sites and cause gain/loss of m7G sites simultaneously. For each m7G site-variant pair, the association of them was measured by the association levels as well as the confidence levels. The association level qualifies the influence that variants exert on m7G sites into the range [0,1]. The closer the association level is to 1, the stronger influence that variant exerts on the exact site. Initially, 812 m7G site-variant pairs with high confidence level were first extracted, then ranked according to the association level. Then 741 m7G site-disease pairs were further picked out with association levels higher than 0.8. Meanwhile, the sequence and genomic location information of m7G-variant pairs were collected correspondingly in this dataset. Specifically, it contains the genomic locations, host genes of m7G sites, site-centered 41 bp reference sequences as well as site-centered 41 bp alternative sequences.

m7G-disease association dataset

In the m7G-disease association dataset, 741 m7G sites were associated with 177 diseases via 741 variants in the m7G-variant dataset. Specifically, these variants are both m7G-associated and disease-associated. In other words, they cause the gain/loss of the m7G site and involve in various disease pathogenesis. Taking these variants as linkages, 177 diseases in ClinVar and GWAS were found to be associated with 741 variants, with implications for disease pathogenesis in m7G RNA methylation.

Methods

m7G-disease association network reconstruction can be transformed into predicting the unknown entries in the m7G-disease association matrix, which can be solved by traditional matrix decomposition methods. However, the number of known associations is so small that matrix decomposition methods cannot achieve satisfactory performance in this case. Thus, we proposed a heterogeneous network-based m7G-disease association prediction method m7GDisAI which will be detailed in the next. The framework of m7GDisAI is shown in Fig. 1.
m⁷G-Disease Association Network

Based on the m⁷G-disease association dataset, the m⁷G-disease adjacency network was constructed to record their associations. To be more specific, let $S = \{s_1, s_2, ..., s_m\}$ and $D = \{d_1, d_2, ..., d_n\}$ denote $m$ m⁷G sites and $n$ diseases respectively. Let $A_{SD} \in \mathbb{R}^{m \times n}$ indicate the adjacency network, $A_{SD_{ij}}$ is 1 if there exists a validated association between m⁷G-disease pair $(s_i, d_j)$. The m⁷G-disease association matrix $A_{SD}$ was provided in Additional file 4: Table S4.

m⁷G similarity networks

As a kind of auxiliary information, m⁷G similarity information plays a critical role in m⁷G-disease association prediction. To make full advantages of the information of m⁷G sites, a series of m⁷G similarity networks were constructed for further use in the heterogeneous network.

m⁷G chemical similarity network

m⁷G chemical similarity network (CSN) depicts the m⁷G similarities in terms of the chemical properties extracted from m⁷G site-centered sequences [23, 24]. Specifically, either sequence is a combination of four nucleotides A,
T, C, G. Each nucleotide can be characterized by three distinct structural chemical properties, such as ring structures, hydrogen bonds, and functional groups. In terms of ring structures, A and G have two benzene rings, while C and T have only one. As for the number of hydrogen bonds formed during hybridization, A and T have two, while G and C have three. Regarding the functional groups they contain, A and C contain amino groups, whereas G and T contain keto groups. Therefore, the $i$-th nucleotide in sequence $N$ can be encoded by a vector $(x_i, y_i, z_i)$.

\[
x_i = \begin{cases} 
1 & \text{if } N_i \in \{A, G\} \\
0 & \text{if } N_i \in \{C, T\}
\end{cases}, \quad 
y_i = \begin{cases} 
1 & \text{if } N_i \in \{A, T\} \\
0 & \text{if } N_i \in \{G, C\}
\end{cases}, \quad 
z_i = \begin{cases} 
1 & \text{if } N_i \in \{A, C\} \\
0 & \text{if } N_i \in \{G, T\}
\end{cases}
\]

Therefore, A, C, G, T can be encoded as (1,1,1), (0,0,1), (1,0,0) and (0,1,0) respectively. Thus, the chemical feature of site $s_i$, denoted as $\text{CF}(s_i)$, is the combination of these four vectors, in the form of a sequence consisting of \{0,1\}. Considering the binary numerical properties of the $m7G$ chemical features, the Jaccard coefficient was applied to them. To be specific, for two sites $s_i$ and $s_j$, their pairwise chemical similarity is defined as (1).

\[
\text{che}_\text{sim}_{ij} = \frac{|\text{CF}(s_i) \cap \text{CF}(s_j)|}{|\text{CF}(s_i) \cup \text{CF}(s_j)|}
\]

Then in the $m7G$ CSN, $s_1, s_2, \ldots, s_m$ are nodes, and the edges between them are weighted by the pairwise chemical similarity above. For convenience, the adjacency matrix was indicated as $A_{\text{CSN}}$ (Additional file 5: Table S5).

$m7G$ Cumulative Nucleotide Frequency Similarity Network Similar to the construction of CSN, $m7G$ cumulative nucleotide frequency (CNF) features were extracted for further similarity calculation. To be specific, CNF of the $i$-th nucleotide in a sequence is defined as the sum of all the instances of this nucleotide before the $i+1$ position dividing $i$. Taking the sequence ‘TAA GTCCA’ as an example, the CNF for A is 0.5(1/2), 0.667(2/3), 0.375(3/8) at the 2nd, 3rd and 8th positions respectively. Thus, the CNF features of site $s_i$ are denoted as $\text{CNF}(s_i)$. Comparing with the $m7G$ chemical features, CNF features pay more attention to the sequence context around the $m7G$ site. Then the Cosine coefficient was adopted to calculate similarities of CNF since it reflects the similarity in trend rather than absolute values. For sites $s_i$ and $s_j$, the pairwise CNF similarity is defined as (2).

\[
\text{CNF}_\text{sim}_{ij} = \frac{|\text{CNF}(s_i) \cdot \text{CNF}(s_j)|}{||\text{CNF}(s_i)||_2||\text{CNF}(s_j)||_2}
\]

Then $m7G$ CNF similarity network (CNFSN) was obtained with the weights between nodes $s_i$ and $s_j$ ($i=1,2,\ldots,m$, $j=1,2,\ldots,m$), and the adjacency matrix was indicated as $A_{\text{CNFSN}}$ (Additional file 6: Table S6).

$m7G$ integrated similarity network Since $m7G$ chemical similarity and CNF similarity measure $m7G$ similarities from their own views, we took a linear combination of those two similarities to form an integrated similarity, and the contribution of $m7G$ chemical similarity and CNF similarity is weighted by $\alpha$. For sites $s_i$ and $s_j$, the integrated similarity is defined as (3).
The value of $\alpha$ was chosen from 0 to 1 with step 0.1, and was determined by tenfold cross validation experiments. Then a series of m$^7$G integrated similarity networks were obtained via taking (3) as weights between nodes $s_i$ and $s_j$, ($i=1,2...m$, $j=1,2...m$), and its adjacency matrix was indicated as $A_{SS}$. In addition, if $\alpha$ is 0, then $A_{SS}$ is $A_{CSN}$ while if $\alpha$ is 1, then $A_{SS}$ is $A_{CNFSN}$.

**Disease semantic similarity network**

Disease semantic similarity network (DSSN), indicated by adjacency matrix $A_{DD}$, was also constructed by calculating pairwise disease semantic similarities. Generally speaking, functional similarity between molecules results in similar phenotypes, such as diseases. Based on this fact, many researchers [15, 25–27] utilized functional similarities of the disease-associated molecules for semantic disease similarities. We followed Wang’s PBPA method, which was implemented to calculate pairwise disease semantic similarities [28, 29]. Additionally, the “DisSetSim” web server can be accessed from [http://www.bio-annotation.cn:18080/DincRNAClient](http://www.bio-annotation.cn:18080/DincRNAClient). By calculating all pairwise semantic similarities in $D$, a disease semantic similarity network was obtained and the adjacency matrix was indicated as $A_{DD}$ (Additional file 7: Table S7).

**m$^7$G-disease heterogeneous network**

The m$^7$G-disease heterogeneous network and its adjacency matrix are shown in Fig. 2. The m$^7$G-disease heterogeneous network was constructed by incorporating m$^7$G-disease adjacency network, disease semantic similarity network DSSN, and m$^7$G integrated similarity networks. It was represented by adjacency matrix $A$ and mask matrix $W$, as (4).

$$
A = \begin{pmatrix}
A_{SS} & A_{SD} \\
A_{TD} & A_{DD}
\end{pmatrix},
W = \begin{pmatrix}
W_{SS} & W_{SD} \\
W_{TD} & W_{DD}
\end{pmatrix}
$$

(4)

where $W_{SS}$ and $W_{DD}$ are all one's matrix. For $W_{SD}$, $W_{ij} = 1$ if the association of the $i$-th site to the $j$-th disease is known, 0, vice versa.
By incorporating DSSN and \( m^7 \)G integrated similarity networks into the \( m^7 \)G-disease adjacency network, cold start issue is avoided, while information of sites and diseases is fully be used.

**\( m^7 \)G-disease association inference based on heterogeneous network**

Based on the \( m^7 \)G-disease heterogeneous network constructed above, the goal of recovering \( A_{SD} \) is transformed into completing \( A \). Underpinned by the fact that similar sites have similar molecular pathways for similar diseases, the matrix completion model assumes that the underlying latent factors determining \( m^7 \)G-disease associations are highly correlated. In addition, if two sites are similar, then they would have similar patterns with any other sites, and it is true for diseases. The number of independent factors that govern the pattern of \( A \) is much smaller than that of sites and diseases. In a mathematical view, the number of independent factors is the rank, here we used \( k \) to denote it. Thus, the goal of completing \( A \) can be achieved by the classical matrix decomposition method, which achieved positive results in many cases and is easy to realize. The primary idea of matrix decomposition is to map the adjacency matrix \( A \) into a \( k \) dimensional space, where \( k < m + n \), so dimension reduction is achieved and a lower-dimensional representation of \( A \) in a \( k \)-dimensional space is given by two matrices \( U \in \mathbb{R}^{(m+n) \times k} \) and \( V \in \mathbb{R}^{(m+n) \times k} \). Then \( A \) can be approximated by (5).

\[
A \approx UV^T \tag{5}
\]

The fundamental idea of finding suitable factor matrices \( U, V \) is to minimize the objective function defined as (6):

\[
\min_{U,V} ||W \odot (A - UV^T)||_F^2 \tag{6}
\]

where \( ||*||_F \) is the Frobenius norm, \( W \odot (A - UV^T) \) denotes the Hadamard product of two matrices \( W \) and \( A - UV^T \).

Furthermore, regularization terms should be considered, and the loss function is defined as (7), while the objective function is (8).

\[
L = ||W \odot (A - UV^T)||_F^2 + \lambda_1 ||U||_F^2 + \lambda_2 ||V||_F^2 \tag{7}
\]

\[
\min_{U,V} L \tag{8}
\]

where \( \lambda_1 ||U||_F^2 + \lambda_2 ||V||_F^2 \) is the regularization term to avoid overfitting, with \( \lambda_1 \) and \( \lambda_2 \) being the regularization parameters.

\( \lambda_1 \) and \( \lambda_2 \), which were optimized by cross validation, help to achieve the trade-off between fitting and generalization. The Alternating Least Square method [30, 31] was then followed to reach the global minimum concerning to \( U \) and \( V \). Finally, unknown entries in \( A_{SD} \) were predicted. The implementation process of \( m^7 \)GDisAI is given below.
Results

Experimental design

To systematically evaluate the prediction performance of m\(^7\)GDisAI on the m\(^7\)G-disease association dataset, tenfold cross validation and LOOCV strategies were adopted for the experiments.

As for tenfold cross validation, in the m\(^7\)G-disease association dataset, there are 768 validated known associations, and the others that haven’t been validated are considered as candidate associations. All known associations are randomly divided into 10 sets that are roughly equal size. Each set is taken as test set in turn, in other words, pretends to be unknown ones, while the remaining nine sets serve as the training set. After performing m\(^7\)GDisAI on training set, the test associations were ranked together with the candidate associations in descending order according to the predicted value obtained.
by m\(^7\)GDisAI. Additionally, two types of LOOCV, global LOOCV and local LOOCV, were further carried out on the m\(^7\)G-disease association dataset. At each iteration, each validated known m\(^7\)G-disease association was treated as the test data and all the remaining associations as the training data. The only difference between them is the selection of candidate samples. To be specific, in global LOOCV, the candidate samples are all unknown m\(^7\)G-disease associations, while in local LOOCV, candidate samples are only those associations under the disease of interest. In each scheme of LOOCV, the test sample was ranked with candidate samples in descending order.

Regardless of tenfold cross validation, global LOOCV and local LOOCV, for a given threshold \(\tau\), a test association is regarded as true positive (TP) if it ranks above the threshold, false negative (FN) otherwise. Similarly, a candidate sample is considered as false positive (FP) if it ranks above the threshold, true negative (TN) otherwise. By varying \(\tau\), true positive rate (TPR), false positive rate (FPR) can be calculated for Receiver Operating Characteristic (ROC) curve. It depicts the relative tradeoffs of prediction performance between TP and FP [32]. The area under ROC curve (AUC), ranging from 0 to 1, can be used to evaluate the overall performance [32, 33].

### Parameter setting

There are four parameters, rank \(k\), linear combination coefficient \(\alpha\), regularization parameters \(\lambda_1\) and \(\lambda_2\), that are required to be optimized to enhance the performance of m\(^7\)GDisAI. To be specific, \(k\) is the number of independent factors that govern the pattern of the heterogeneous matrix \(A\), and if \(k\) is too large, then the algorithm would be time-consuming. Then \(k\) is chosen from \{70,90,110\}. The linear combination coefficient \(\alpha\) weights the contribution of m\(^7\)G chemical similarity and m\(^7\)G CNF similarity in m\(^7\)G integrated similarity network, and it was taken from 0 to 1.0 with the step 0.1. In addition, regularization parameters \(\lambda_1\) and \(\lambda_2\) control the relative penalty extent of the factor matrices \(U\) and \(V\) respectively, and they were chosen from \{2\(^{-2}\), 2\(^{-1}\), 2\(^0\), 2\(^1\), 2\(^2\)\}. It is apparent that \(k, \lambda_1\) and \(\lambda_2\) directly influence the optimal solution of the two factor matrices \(U\) and \(V\), while \(\alpha\) only has an impact on the m\(^7\)G similarity matrix \(A_{SS}\). Thus, \(\alpha\) was first fixed to 0.5 or any other specific value between 0 to 1, and a grid search strategy was performed on \(k, \lambda_1\) and \(\lambda_2\). tenfold cross validation experiments were performed with all combination of \(k, \lambda_1\) and \(\lambda_2\) on the training set. m\(^7\)GDisAI performed best when \(k\) is 90, \(\lambda_1\) is -2 and \(\lambda_2\) is -2 with AUC of 0.728. For fairness, the impact of \(\alpha\) on m\(^7\)GDisAI was measured via tenfold cross validation experiments with fixed \(k, \lambda_1\) and \(\lambda_2\). To be specific, \(\alpha\) is 0 means that \(A_{SS}\) is \(A_{CHN}\) and m\(^7\)GDisAI only utilizes m\(^7\)G chemical similarities, while \(\alpha\) is 1 indicates that \(A_{SS}\) is \(A_{CNFHN}\) and m\(^7\)GDisAI only utilizes m\(^7\)G CNF similarities. Table 1 reports the AUC scores with all \(\alpha\), and the highest AUC score is marked in bold.

| \(\alpha\) | 0    | 0.1  | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.8  | 0.9  | 1.0  |
|-----------|------|------|------|------|------|------|------|------|------|------|------|
| AUC       | 0.700| 0.703| 0.706| 0.722| 0.705| 0.728| 0.731| 0.733| 0.737| 0.740| 0.742|

Table 1: AUC scores of different \(\alpha\) in the tenfold cross validation experiments.
In Table 1, as \( \alpha \) increases, AUC scores generally show an increased tendency except when \( \alpha \) is 0.4, and reaches its maximum at 0.742 when \( \alpha \) is 1. In other words, the more CNF similarities contribute, the higher the AUC scores achieved, and \( m^7 \)GDisAI has the best performance when only utilizes CNFH. Table 1 validates the effectiveness of the CNF features and Cosine coefficient to some extent. Specifically, chemical features decode the nucleotides of \( m^7 \)G site-centered sequence individually, while CNF features pay more attention to the context of site-centered sequence. Meanwhile, the Cosine coefficient reflects the similarity in trend instead of absolute value as the Jaccard coefficient calculates.

**Performance evaluation**

To further evaluate the robustness of \( m^7 \)GDisAI, we conducted 10 runs of tenfold cross validation experiments by taking \( \alpha \) as 1, which has the best performance in the Table 1. The mean value of AUC scores is 0.740 with standard variance at 0.0024, showing the effectiveness and stability of \( m^7 \)GDisAI. Figure 3a clearly displays the ROC curves with respect to the best performance in tenfold cross validation experiments. Additionally, LOOCV experiments were further conducted to comprehensively evaluate the performance of \( m^7 \)GDisAI. The AUC of global LOOCV was 0.769 while that of local LOOCV was 0.635. The ROC curves of LOOCV experiments are illustrated in the Fig. 3b.

As we can see from Fig. 3b, local LOOCV experiment performs worse than global LOOCV. The key factor contributing to this phenomenon is the number of candidate samples that the test sample were ranked with. To be specific, the number of candidate samples participating in global LOOCV is much larger than those involved in the local LOOCV. In other words, the local LOOCV experiments have more rigorous requirements for positive results.

![Fig. 3](image_url) The best performance of \( m^7 \)GDisAI for tenfold cross validation and LOOCV experiments. **a.** ROC curves generated by tenfold cross validation. **b.** ROC curves generated by global LOOCV and local LOOCV.
Case study

Ovarian cancer is the most common cause of gynecological cancer-associated death [34]. Over the past decades, the overall cure rate remains approximately 30% [35]. The reason for low cure rate is the late presentation in most cases. 80% of patients have symptoms, however, these symptoms are shared with many more common gynecological conditions [35]. Given the heterogeneity of this disease, it is necessary to explore the disease pathogenesis at molecular and cellular levels. Then taking all known associations as training samples, while other unknown ones as candidate samples. Since CNFHN has the best performance in the tenfold cross validation experiments, then we performed it on the training samples to score the candidate samples, especially those under ovarian cancer. Furthermore, all the m7G sites were ranked in descending order according to their association scores with ovarian cancer, and the top 100 m7G sites were selected as potential ovarian cancer-associated sites. 98 host genes of these sites were further mapped out. To predict potential cellular processes and molecular functions that involve m7G methylation, we used the R package “clusterProfiler” to analyze and visualize the functional profiles of m7G host genes.

GO terms include three subontologies, cellular component (CC), biological process (BP) and molecular function (MF), and they can be conducted via enrichGO function. In the parameter setting of the enrichGO function, we set the parameter “ont” to “ALL”, aiming at performing CC, BP and MF together. Additionally, the \( p \)-value cutoff was set as 0.05, \( q \)-value cutoff 0.2, indicating statistical significance of associations between host genes and GO terms. Furthermore, “BH” method was used to adjust the \( p \)-value to control the false discovery rate, which was considered to be statistically significant. Considering the potentially biological complexities in which a gene may belong to multiple annotation categories, we utilized a gene-concept network to depict the linkages of gene and GO terms as a network. Figure 4 provides a visualization of the gene-concept network by cnetplot function.

In Fig. 4, ten most significantly enriched terms including CC, BP and MF were shown to be associated with 26 genes. The enrichment analysis results have been verified by

![Fig. 4](image_url)

**Fig. 4** The gene-concept network of functional GO enrichment results. The connection between a gene and a term means that the gene is involved in this GO term.
published literature. Specifically, TP53 is the most widely studied tumor suppressor gene [36], and it is the host gene of m7G_ID_194615, m7G_ID_203640, m7G_ID_202781, m7G_ID_194736 and m7G_ID_280795 as Additional file 1: Table S1 shows. TP53 functions in ovarian cancer by arresting the cell cycle at G1 phase and by triggering apoptosis [37]. In addition, Lang et al. [38] found that UV radiation leads to base-pair changes of p53, the protein product of the TP53 gene, and further leads to tumor formation. Furthermore, Jeremy et al. [39] experimentally showed that the dynamic patterns of TP53 vary depending on the stimulus. For example, the levels of p53 exhibit a series of pulses with fixed amplitude and frequency in response to DNA breaks caused by γ-irradiation. These discoveries prove that TP53 is enriched into “negative regulation of mitotic cell cycle”, “response to UV” and “cellular response to environmental stimulus” terms [40].

To date, hereditary nonpolyposis colorectal cancer (HNPCC) is the third major cause of hereditary ovarian cancer, and HNPCC is caused by mutations in genes involved in DNA mismatch repair [41]. MLH1 [42] (host gene of m7G_ID_137019, m7G_ID_137020, m7G_ID_151088, m7G_ID_220822), MSH2 [43] (host gene of m7G_ID_161433, m7G_ID_192868, m7G_ID_253317), MSH6 [44] (host gene of m7G_ID_200227, m7G_ID_317794) and PMS2 [45] (host gene of m7G_ID_155289) are all reported to be mismatch repair genes. To be specific, the MLH1 and MSH2 genes are the most common genes for HNPCC-associated ovarian cancer, and account for 80%-90% of observed mutations [46]. What’s more, Cederquist et al. [47] reported that ovarian cancer is in the MSH6 tumor spectrums. Besides, PIK3CA was also known to be oncoproteins of ovarian cancer [48], and they are the host genes of m7G_ID_2249, m7G_ID_9238 in Additional file 1: Table S1 respectively. Notably, PIK3CA activated mutation participates in the PI3K pathway which is activated in approximately 70% of ovarian cancer [49], and is enriched in regulation of protein kinase B signaling, which is activated by autocrine or paracrine signaling through protein kinase signaling in many kinds of cancers [49].

Numerical cases [50–52] have suggested that ERBB family of receptor tyrosine kinases has a significant contribution to the initiation and progression of ovarian cancer. EGFR and ERBB2 in Fig. 4 are members of the ERBB family of receptor tyrosine kinases. EGFR is the host gene of m7G_ID_149119 and its overexpression has been observed in 30%-98% of epithelial ovarian cancer in all histologic subtypes, and enhanced expression of EGFR is correlated with advanced-stage disease as well as poor response to chemotherapies. Additionally, Ginath et.al reported [53] that ERBB2 (host gene of m7G_ID_268139) activates multiple downstream signaling pathways, and then promotes the proliferation, invasion, and metastasis of tumor cells.

**Discussion**

This research into identifying potential m7G-disease association prediction will help us understand the pathogenesis of diseases and promote the treatment of diseases. In this paper, we extracted 768 associations between 741 m7G sites and 177 diseases to construct the m7G-disease association dataset. To predict the m7G-disease association based on the m7G-disease dataset, we proposed a heterogeneous network-based association inference method m7GDisAI. For m7GDisAI, we performed m7G-disease
association inference on a series of heterogeneous networks which contain m7G-disease adjacency network and disease semantic similarity network, but different m7G similarity networks, CHN, CNFHN and their combinations. 10-fold cross validation, global and local LOOCV were performed with m7GDisAI. CNFHN outperforms the CHN and other heterogeneous networks, which proves the effectiveness of CNF features. Then a case study of ovarian cancer was later conducted by CNFHN. It is worth mentioning that the constructed m7G-variant pair dataset and m7G-disease association dataset may play important role in further investigation of disease-associated m7G sites discovery. To our knowledge, m7GDisAI is the first algorithm that connects m7G sites, variants as well as diseases together to uncover potential cancer-related functions of m7G, which may provide some valuable hints for wet experiments guidance. However, there remains limitations in this study. Firstly, the research of m7G and diseases is an ongoing topic and the m7G-disease dataset is far from completed. Secondly, more feature selection methods could be taken into consideration to construct m7G similarity networks and further improve the accuracy of m7GDisAI.

Conclusions
m7GDisAI is a heterogeneous network-based m7G-disease association inference method and is freely accessible at http://180.208.58.66/m7GDisAI/. m7GDisAI uncovers disease-associated m7G sites by applying matrix decomposition method on a heterogeneous network-based m7G-disease association matrix. m7GDisAI provides users a function to query related m7G sites of disease which the users are interested in. The website hosts the top 20 m7G sites predicted to be associated with 177 diseases with high prediction scores, which may provide some clues for pathogenesis of diseases. The front-end is implemented in JavaScript while the back-end is implemented in Python as well as R. We will continue updating m7GDisAI by adding additional information, improving the implementation, and incorporating new measures for inferring disease-associated m7G sites. The user can always access the latest version of m7GDisAI.

Availability and requirements
Project name: m7GDisAI. Project home page: http://180.208.58.66/m7GDisAI/. Operating system(s): Linux, Windows. Programming language: Python, R, JavaScript. Other requirements: Not specified. Python version 3.8.0 or higher, R version 4.0.3 or higher. License: GNU GPL. Any restrictions to use by non-academics: None.

Abbreviations
m7G: N7-methylguanosine; m7G-MeRIP-Seq: N7-methylguanosine Methylated RNA immunoprecipitation sequencing; m7G-miCLIP-Seq: N7-methylguanosine Individual-Nucleotide-Resolution Crosslinking and Immunoprecipitation; m7GDisAI: N7-methylguanosine-disease association inference; CHN: Chemical Heterogeneous Network; CNF: Cumulative Nucleotide Frequency; CNFHN: Cumulative Nucleotide Frequency Heterogeneous Network; LOOCV: Leave-one-out cross validation; DSSN: Disease Semantic Similarity Network; CSN: Chemical Similarity Network; CNFSN: Cumulative Nucleotide Frequency Similarity Network; ISN: Integrated Similarity Network; MICA: Most Informative Common Ancestor; ALS: Alternating Least Squares; FP: False Positive; TN: True Negative; FN: False Negative; ROC: Receiver Operating Characteristic Curves.
Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04007-9.

Acknowledgements
Not applicable.

Authors’ contributions
JM and LZ built the architecture for m7GDisAI, designed and implemented the experiments, analyzed the result, and wrote the paper. JC analyzed the result, and revised the paper. BS prepared the data. CZ built up the webserver. HL supervised the project, analyzed the result, and revised the paper. All authors read and approved the final manuscript.

Funding
This work has been supported by Postgraduate Student Education Reform Research and Practice Funds (Research Projects No. 2019YJSJG045 to LZ), the National Natural Science Foundation of China (Research Projects Nos. 61971422 to LZ; 31871337 to HL). The funding body did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The detailed information of m7G-variant dataset is listed in Additional file 1: Table S1. For each m7G-disease pair, information for their sequence and genomic location is included. Additional file 2: Table S2 shows diseases we collected with their names and DOID. Additional file 3: Table S3 provides the information for m7G-disease association dataset with 768 known m7G-disease associations. In addition, Additional file 4: Table S4 is the m7G-disease matrix ASD where the validated associations are all one. Additional files 5: Table S5–Additional file 6: Table S6 are m7G similarity networks ACSN, ACNFSN respectively, while Additional file 7: Table S7 is the disease semantic similarity network ADD. Furthermore, Additional file 8: Table S8 presents the recommended m7G sites and their host gene of ovarian cancer. The website m7GDisAI implemented to query related m7G sites of the disease which you are interested in is deposited at http://180.208.58.66/m7GDisAI/.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
1 Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China. 2 School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China. 3 Department of Biological Sciences, AI University Research Center, Xi’an Jiaotong-Liverpool University, Suzhou 215123, China.

Received: 28 June 2020   Accepted: 8 February 2021
Published online: 24 March 2021

References
1. Jaffrey SR. An expanding universe of mRNA modifications. Nat Struct Mol Biol. 2014;21(11):945–6.
2. Zaccara S, Ries RJ, Jaffrey SR. Reading, writing and erasing mRNA methylation. Nat Rev Mol Cell Biol. 2019;20(10):608–24.
3. Guy MP, Phizicky EM. Two-subunit enzymes involved in eukaryotic post-transcriptional rRNA modification. RNA Biol. 2014;11(12):1608–18.
4. Sibani KE, Warda AS, Sharma S, Entian K-D, Lafontaine DLJ, Bohnscck MT. Tuning the ribosome: the influence of rRNA modification on eukaryotic ribosome biogenesis and function. RNA Biol. 2017;14(9):1138–52.
5. Cowlng VH. Regulation of mRNA cap methylation. Biochemical Journal. 2010;435:295–302.
6. Malbec L, Zhang T, Chen Y-S, Zhang Y, Sun B-F, Shi B-Y, Zhao Y-L, Yang Y, Yang Y-G. Dynamic methylome of internal mRNA N-7-methylguanosine and its regulatory role in translation. Cell Res. 2019;29(11):927–41.
7. Furuichi Y. Discovery of m(7)G-cap in eukaryotic mRNAs. Proc Jpn Acad Ser B-Phys Biol Sci. 2015;91(8):394–409.
8. Shi HJ, Moore PB. The crystal structure of yeast phenylalanine tRNA at 1.93 angstrom resolution: A classic structure revisited. RNA. 2000, 6(8):1091–1105.
9. Oliva R, Cavallo L, Tramontano A. Accurate energies of hydrogen bonded nucleic acid base pairs and triplets in RNA tertiary interactions. Nucleic Acids Res. 2006;34(3):865–79.
10. Shimotomino K, Kodama Y, Hashimoto J, Miura KI. Importance of 5’-terminal blocking structure to stabilize mRNA in eukaryotic protein synthesis. Proc Natl Acad Sci USA. 1977;74(7):2734–8.
11. Pei Y, Shuman S. Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5. J Biol Chem. 2002;277(22):19639–48.
12. Konarska MM, Padgett RA, Sharp PA. Recognition of cap structure in splicing in vitro of mRNA precursors. Cell. 1984;38(3):731–6.
13. Drummond DR, Armstrong J, Colman A. The effect of capping and polyadenylation on the stability, movement and translation of synthetic messenger RNAs in Xenopus oocytes. Nucleic Acids Res. 1985;13(20):7375–94.
14. Lewis JD, Izaurralde E. The role of the cap structure in RNA processing and nuclear export. Eur J Biochem. 1997;247(2):461–9.
15. Muthukrishnan S, Both GW, Furuichi Y, Shatkin AJ. 5’-Terminal 7-methylguanosine in eukaryotic mRNA is required for translation. Nature. 1975;255(5503):33–7.
16. Shaheen R, Abdel-Salam GMH, Guy MP, Alomar R, Abdel-Hamid MS, Afsh HH, Ismail SL, Emam BA, Phizicky EM, Alkuraya FS. Mutation in WDR4 impairs tRNA m(7)G(46) methylation and causes a distinct form of microcephalic primordial dwarfism. Genome Biol. 2015, 16.
17. Trimouille A, Lasseaux E, Barat P, Deiller C, Drunat S, Rooryck C, Arveiler B, Lacombe D. Further delineation of the phenotype caused by biallelic variants in the WDR4 gene. Clin Genet. 2018;93(2):374–7.
18. Lin S, Liu Q, Lelyveld VS, Choe J, Szostak JW, Gregory RI. Mettl1/Wdr4-mediated m(7)G tRNA methylation is required for normal mRNA translation and embryonic stem cell self-renewal and differentiation. Mol Cell 2018, 71(2):244–55.
19. Pereira PL, Magnol L, Sahin I, Brault V, Duchon A, Prandini P, Grout A, Bizot J-C, Chadeaux-Vekemans B, Deutsch S, et al. A new mouse model for the trisomy of the Abcg1-U2af1 region reveals the complexity of the combinatorial genetic code of down syndrome. Hum Mol Genet. 2009;18(24):4756–69.
20. Barbieri I, Tzelepis K, Pandolfini L, Shi J, Millan-Zambrano G, Robson SC, Aspiris D, Migliori V, Bannister AJ, Han N et al. Promoter-bound METTL3 maintains myeloid leukaemia by m(6)A-dependent translation control. Nature 2017, 552(7683):126–9.
21. Zhang LS, Liu C, Ma HH, Dai Q, Sun HL, Luo GZ, Zang JG, Zhang LD, Hu LL, Dong XY et al. Transcriptome-wide mapping of internal N-7-methylguanosine methylome in mammalian mRNA. Mol Cell 2019, 74(6):1304.
22. Song B, Tang Y, Chen K, Wei Z, Rong R, Lu Z, Su J, de Magalhaes JP, Rigden DJ, Meng J. m7GHub: deciphering the location, regulation and pathogenesis of internal m7G-methylguanosine (m7G) sites in human. Bioinformatics (Oxford, England) 2020.
23. Chen K, Wei Z, Zhang Q, Wu X, Rong R, Lu Z, Su J, de Magalhaes JP, Rigden DJ, Meng J. WHISTLE: a high-accuracy map of the human N6-methyladenosine (m6A) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res 2019, 47(7).
24. Zhou Y, Zeng P, Li Y-H, Zhang Z, Cui Q. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features. Nucleic Acids Res 2016, 44(10).
25. Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. J Biomed Inform. 2012;45(2):363–71.
26. Cheng L, Li J, Ju P, Peng J, Wang Y. SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. Plos One 2014, 9(6).
27. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. 1995.
28. Wang JZ, Du Z, Payatakeel R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81.
29. Hu Y, Zhao L, Liu Z, Ju H, Shi H, Xu P, Wang Y, Cheng L. DisSetSim: an online system for calculating similarity between disease sets. J Biomed Semant. 2017; 8.
30. Jain P, Netrapalli P, Sanghavi S, et al. A new mouse model for the trisomy of the Abcg1-U2af1 region reveals the complexity of the combinatorial genetic code of down syndrome. Hum Mol Genet. 2009;18(24):4756–69.
31. Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. J Biomed Inform. 2012;45(2):363–71.
32. Cheng L, Li J, Ju P, Peng J, Wang Y. SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. Plos One 2014, 9(6).
33. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. 1995.
34. Wang JZ, Du Z, Payatakeel R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81.
35. Hu Y, Zhao L, Liu Z, Ju H, Shi H, Xu P, Wang Y, Cheng L. DisSetSim: an online system for calculating similarity between disease sets. J Biomed Semant. 2017; 8.
36. Jain P, Netrapalli P, Sanghavi S, et al. A new mouse model for the trisomy of the Abcg1-U2af1 region reveals the complexity of the combinatorial genetic code of down syndrome. Hum Mol Genet. 2009;18(24):4756–69.
37. Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. J Biomed Inform. 2012;45(2):363–71.
38. Cheng L, Li J, Ju P, Peng J, Wang Y. SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. Plos One 2014, 9(6).
39. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. 1995.
40. Wang JZ, Du Z, Payatakeel R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81.
41. Hu Y, Zhao L, Liu Z, Ju H, Shi H, Xu P, Wang Y, Cheng L. DisSetSim: an online system for calculating similarity between disease sets. J Biomed Semant. 2017; 8.
42. Jain P, Netrapalli P, Sanghavi S, et al. A new mouse model for the trisomy of the Abcg1-U2af1 region reveals the complexity of the combinatorial genetic code of down syndrome. Hum Mol Genet. 2009;18(24):4756–69.
40. Batchelor E, Loewer A, Mock C, Lahav G: Stimulus-dependent dynamics of p53 in single cells. Mol Syst Biol. 2011, 7.
41. Pal T, Permutt-Wey J, Sellers TA. A review of the clinical relevance of mismatch-repair deficiency in ovarian cancer. Cancer. 2008,113(4):733–42.
42. Bronner CE, Baker SM, Morrison PT, Warren G, Smith LG, Lescoc MK, Kane M, Earabino C, Lipford J, Lindblom A. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature. 1994,368(6468):258–61.
43. Samimi G, Fink D, Varki NM, Husain A, Hoskins WJ, Alberts DS, Howell SB. Analysis of MLH1 and MSH2 expression in ovarian cancer before and after platinum drug-based chemotherapy. Clin Cancer Res. 2000,6(4):1415–21.
44. Miyaki M, Konishi M, Tanaka K, Kikuchi-Yanoshita R, Muraoka M, Yasuno M, Igar T, Koike M, Chiba M, Mori T. Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. Nat Genet. 1997,17(3):271–2.
45. Lum CT. Sun RW-Y, Zou T, Che C-M: Gold(III) complexes inhibit growth of cisplatin-resistant ovarian cancer in association with upregulation of proapoptotic PMS2 gene. Chem Sci. 2014,5(4):1579–84.
46. Ichikawa Y, Lemon SJ, Wang S, Franklin B, Watson P, Knezetic JA, Bewtra C, Lynch HT. Microsatellite instability and expression of MLH1 and MSH2 in normal and malignant endometrial and ovarian epithelium in hereditary nonpolyposis colorectal cancer family members. Cancer Genet Cytogenet. 1999,112(1):2–8.
47. Cederquist K, Emanuelsson M, Wiklund F, Golovleva I, Palmaqvist R, Gronberg H. Two Swedish founder MSH6 mutations, one nonsense and one missense, conferring high cumulative risk of Lynch syndrome. Clin Genet. 2005,68(6):533–41.
48. Shayesteh L, Lu Y, Kuo WL, Baldocchi R, Godfrey T, Collins C, Pinkel D, Powell B, Mills GB, Gray J.W. PIK3CA is implicated as an oncogene in ovarian cancer. Nat Genet. 1999,21(1):99–102.
49. Lee S, Choi EJ, Jin CB, Kim DH. Activation of PI3K/Akt pathway by PTEN reduction and PIK3CA mRNA amplification contributes to cisplatin resistance in an ovarian cancer cell line. Gynecol Oncol. 2005,97(1):26–34.
50. Arteaga CL, Engelman JA. ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. Cancer Cell. 2014,25(3):282–303.
51. Riese DJ 2nd, Stern DF. Specificity within the EGF family/ErbB receptor family signaling network. BioEssays. 1998,20(1):41–8.
52. Roskoski R Jr. The ErbB/HER family of protein-tyrosine kinases and cancer. Pharmacol Res. 2014,79:34–74.
53. Ginath S, Menczer J, Friedmann Y, Angorn H, Aviv A, Tajima K, Dantes A, Glezerman M, Vlodavsky I, Amsterdam A. Expression of heparanase, Mdm2, and erbB2 in ovarian cancer. Int J Oncl. 2001,18(6):1133–44.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.