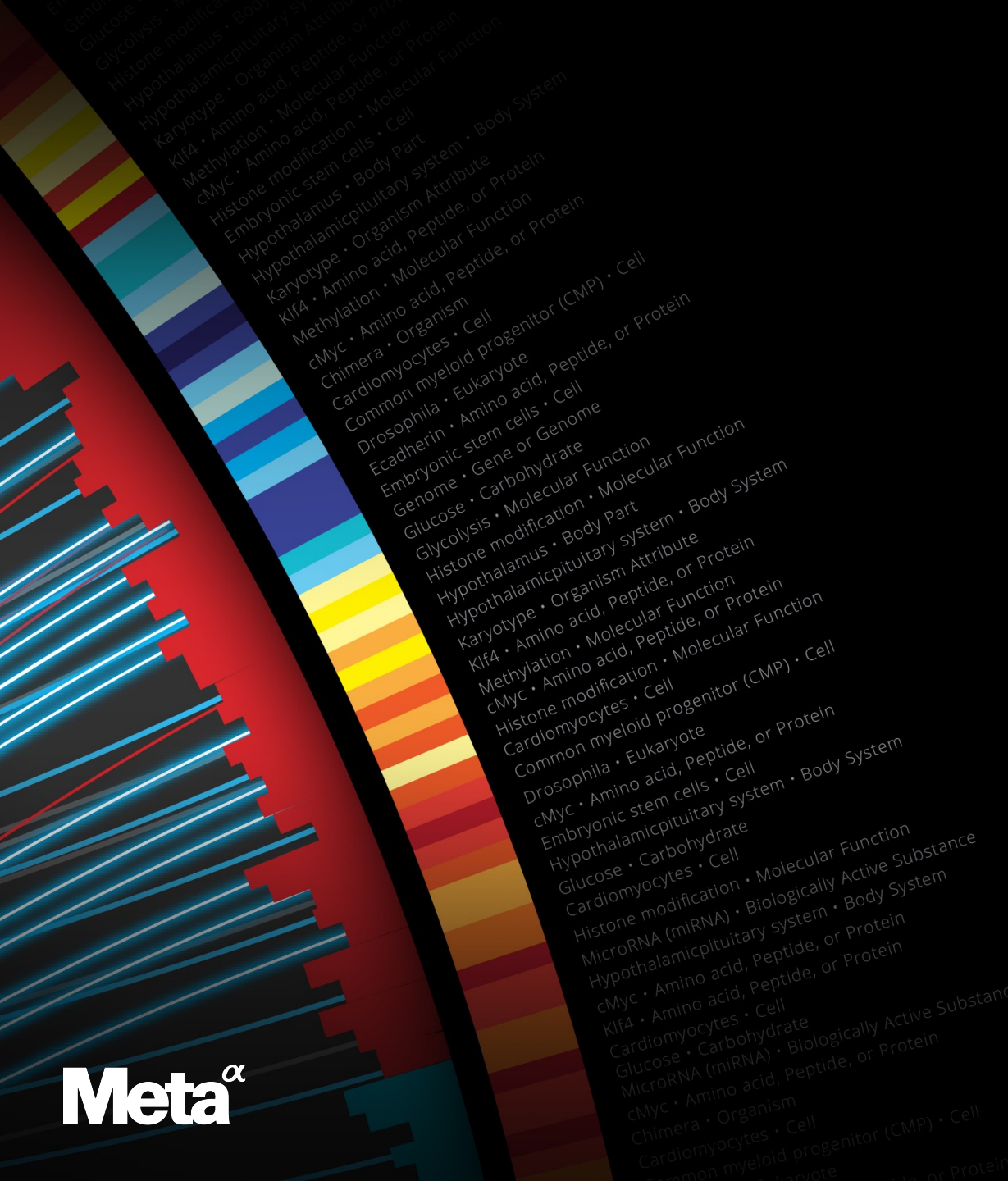# Bibliometric Intelligence

Elizabeth Caley

COO

ec@meta.com

Choi et al.
Functional impacts of amino acid variants
Page 1

Abstract

As next generation sequencing projects generate massive genome-wide sequence variation data, bioinformatics tools are being developed to provide computational predictions on the functional effects of sequence variations and narrow down the search of casual variants for disease phenotypes. Different classes of sequence variations at the nucleotide level are involved in human diseases, including substitutions, insertions, deletions, frameshifts, and non-sense mutations. Frameshifts and non-sense mutations are likely to cause a negative effect on protein function. Existing prediction tools primarily focus on studying the deleterious effects of single amino acid substitutions through examining amino acid conservation at the position of interest among related sequences, an approach that is not directly applicable to insertions or deletions. Here, we introduce a versatile alignment-based score as a new metric to predict the damaging effects of variations not limited to single amino acid substitutions but also in-frame insertions, deletions and multiple amino acid substitutions. This alignment-based score measures the change in sequence similarity of a query sequence to a protein sequence homolog before and after the introduction of an amino acid variation to the query sequence. Our results showed that the scoring scheme performs well in separating disease-associated variants (n=21,662) from common polymorphisms (n=37,022) for UniProt human protein variations, and also in separating deleterious variants (n=15,179) from neutral variants (n=17,891) for UniProt non-human protein variations. In our approach, the area under the receiver operating characteristic curve (AUC) for the human and non-human protein variation datasets is ~0.85. We also observed that the alignment-based score correlates with the deleteriousness of a sequence variation. In summary, we have developed a new algorithm, PROVEAN (**Pro**tein **V**ariation **E**ffect **An**alyzer), which provides a generalized approach to predict the functional effects of protein sequence variations including single or multiple amino acid substitutions, and

What's the topic?

How well does the topic fit my journal?

Who authored it?

What is their publication history?

Which institutes?

Are the findings novel?

Is the science is solid?

How does it rank against other submissions?

Is this a highly citable paper?

Choi et al.
Functional impacts of amino acid variants
Page 1

Abstract

As next generation sequencing projects generate massive genome-wide sequence variation data, bioinformatics tools are being developed to provide computational predictions on the functional effects of sequence variations and narrow down the search of casual variants for disease phenotypes. Different classes of sequence variations at the nucleotide level are involved in human diseases, including substitutions, insertions, deletions, frameshifts, and non-sense mutations. Frameshifts and non-sense mutations are likely to cause a negative effect on protein function. Existing prediction tools primarily focus on studying the deleterious effects of single amino acid substitutions through examining amino acid conservation at the position of interest among related sequences, an approach that is not directly applicable to insertions or deletions. Here, we introduce a versatile alignment-based score as a new metric to predict the damaging effects of variations not limited to single amino acid substitutions but also in-frame insertions, deletions and multiple amino acid substitutions. This alignment-based score measures the change in sequence similarity of a query sequence to a protein sequence homolog before and after the introduction of an amino acid variation to the query sequence. Our results showed that the scoring scheme performs well in separating disease-associated variants (n=21,662) from common polymorphisms (n=37,022) for UniProt human protein variations, and also in separating deleterious variants (n=15,179) from neutral variants (n=17,891) for UniProt non-human protein variations. In our approach, the area under the receiver operating characteristic curve (AUC) for the human and non-human protein variation datasets is ~0.85. We also observed that the alignment-based score correlates with the deleteriousness of a sequence variation. In summary, we have developed a new algorithm, PROVEAN (**Pro**tein **V**ariation **E**ffect **An**alyzer), which provides a generalized approach to predict the functional effects of protein sequence variations including single or multiple amino acid substitutions, and

What's the topic?

How well does the topic fit my journal?

Who authored it?

What is their publication history?

Which institutes?

Are the findings novel?

Is the science is solid?

How does it rank against other submissions?

Is this a highly citable paper?

Choi et al.
Functional impacts of amino acid variants
Page 1

Abstract

As next generation sequencing projects generate massive genome-wide sequence variation data, bioinformatics tools are being developed to provide computational predictions on the functional effects of sequence variations and narrow down the search of casual variants for disease phenotypes. Different classes of sequence variations at the nucleotide level are involved in human diseases, including substitutions, insertions, deletions, frameshifts, and non-sense mutations. Frameshifts and non-sense mutations are likely to cause a negative effect on protein function. Existing prediction tools primarily focus on studying the deleterious effects of single amino acid substitutions through examining amino acid conservation at the position of interest among related sequences, an approach that is not directly applicable to insertions or deletions. Here, we introduce a versatile alignment-based score as a new metric to predict the damaging effects of variations not limited to single amino acid substitutions but also in-frame insertions, deletions and multiple amino acid substitutions. This alignment-based score measures the change in sequence similarity of a query sequence to a protein sequence homolog before and after the introduction of an amino acid variation to the query sequence. Our results showed that the scoring scheme performs well in separating disease-associated variants (n=21,662) from common polymorphisms (n=37,022) for UniProt human protein variations, and also in separating deleterious variants (n=15,179) from neutral variants (n=17,891) for UniProt non-human protein variations. In our approach, the area under the receiver operating characteristic curve (AUC) for the human and non-human protein variation datasets is ~0.85. We also observed that the alignment-based score correlates with the deleteriousness of a sequence variation. In summary, we have developed a new algorithm, PROVEAN (**Pro**tein **V**ariation **E**ffect **An**alyzer), which provides a generalized approach to predict the functional effects of protein sequence variations including single or multiple amino acid substitutions, and

Meta

5

Choi et al.
...l impacts of amino acid variants
Page 1

Projected Citation Count

23

Journal Matches

5

Forecasted Impact Score

0.81

Journal Matching Score

81%

Top

10%

Manuscript Ranking

Abstract

As next generation sequencing... ...genome-wide sequence variation data, bioinformatics tools are being developed to provide computational predictions on the functional effects of sequence variations and narrow down the search of casual variants for disease phenotypes. Different classes of sequence variations at the nucleotide level are involved in human diseases, including substitutions, insertions, deletions, frameshifts, and ...sense mutations. Frameshifts and non-sense mutations are likely to cause a negative effect... ...protein function. Existing prediction tools primarily focus on studying the deleterious effects... single amino acid substitutions through examining amino acid conservation at the position of interest among related sequences, an approach that is not directly applicable to insertions or deletions. Here, we introduce a versatile alignment-based score as a new metric to predict the damaging effects of variations not limited to single amino acid substitutions but also in-frame insertions, deletions and multiple amino acid substitutions. This alignment-based score measures the change in sequence similarity of a query sequence to a protein sequence homolog before and after the introduction of an amino acid variation to the query sequence. Our results showed that the scoring scheme performs well in separating disease-associated variants (n=21,662) from common polymorphisms (n=37,022) for UniProt human protein variations, and also in separating deleterious variants (n=15,179) from neutral variants (n=17,891) for UniProt non-human protein variations. In our approach, the area under the receiver operating characteristic curve (AUC) for the human and non-human protein variation datasets is ~0.8... We also observed that the alignment-based score correlates with the deleteriousness of a... sequence variation. In summary, we have developed a new algorithm, PROVEAN (**Pro**tein **V**ariation **E**ffect **An**alyzer), which provides a generalized approach to predict the functional effects of protein sequence variations including single or multiple amino acid substitutions, and...

Meta$^\alpha$

Bibliometric Intelligence

1.1M Biomedical manuscripts analyzed over 3 years

Examined over 210 characteristics

Bibliometric Intelligence Performance

CITATION PROJECTIONS	Pre-publication article with no journal matching	Pre-publication article with journal match
Pearson correlation	130.0%	136.7%
Spearman correlation	63.6%	68.2%

IMPACT PROJECTIONS	Pre-publication article with no journal matching	Pre-publication article with journal match
Pearson correlation	180.8%	184.6%
Spearman correlation	75.0%	86.1%

Meta$^\alpha$

Largest Commercial STM
Text-Mining Collection on Earth

37
Major STM Publishers

38K
Serial Titles
(Books & Journals)

28M
Closed Access
Full-Text Articles

+

14M
Open Access
Full-Text Articles

$850M
In Peer-Reviewed
Research Content

The World's Largest Scientific Knowledge Graph

3.5B
recommendations

1B
paper-concept matches

422M
citations

26M
papers

16M
genetic elements

20M
concepts

14.5M
researchers

2M
antibodies

407K
drugs

427K
institutes

234K
bacteria

96K
diseases

85K
products

36K
journals

4.6B
Knowledge Graph connections

Meta$^\alpha$