**ICMC - Instituto de Ciências Matemáticas e da Computação**
**Introdução à Bioinformática - SCC0271**
Alec Campos Aoki - 15436800

# Análise da Qualidade de Reads de RNA-Seq de *Nephilia clavipes*

## 1 Introdução

### 1.1 Fonte dos Dados

**Fonte**: Babb, P., Lahens, N., Correa-Garhwal, S. et al. The *Nephila clavipes* genome highlights the diversity of spider silk genes and their complex expression. *Nat Genet* **49**, 895–903 (2017). https://doi.org/10.1038/ng.3852

**Códigos de acesso** (SRA):

- Bioprojeto: PRJNA356433;
- Bioamostra: SAMN06132063;
- Experimento: SRX2458127;
- Run: SRR5139362.

**Descrição**: RNA-Seq de *Nephila clavipes*: corpo inteiro de fêmeas adultas.

## 1.2 Sobre o Sequenciamento

O estudo procurou investigar os genes da spidroína, uma família de proteínas estruturais que compõem as sedas de aranhas, construindo o primeiro genoma de uma aranha tecelã, a *Nephila clavipes*. Foram catalogadas 28 spidroínas *Nephila*, representando todos os tipos conhecidos de spidroínas dessa família de aranhas. A caracterização da expressão da spidroína em tipos distintos de glândulas de seda indica que elas podem expressar vários tipos de spidroína com funções diversas.
O conjunto de dados aqui analisado representa sequenciamento de RNA (RNA-Seq) do corpo inteiro de fêmeas adultas da aranha . O RNA-Seq é uma técnica de sequenciamento de nova geração (NGS) utilizada para analisar o transcriptoma, permitindo a identificação e quantificação de genes expressos. No contexto deste estudo, o objetivo foi caracterizar a diversidade de genes de seda de aranha e sua expressão complexa.

Figura 1: *Nephila clavipes* em sua teia. Fonte: https://ismaeljsnature.blogspot.com/2013/08/aranha-de-teia-dourada-clavipes-aranha.html

# 2 Metodologia

## 2.1 Obtenção dos dados

Acesso à sequência via SRA:

```
prefetch SRX2458127
```

```
2025-09-18T17:20:34 prefetch.3.2.1: 1) Resolving '
    SRX2458127'...
2025-09-18T17:20:41 prefetch.3.2.1: Current preference is
     set to retrieve SRA Normalized Format files with
    full base quality scores
2025-09-18T17:20:42 prefetch.3.2.1: 1) Downloading '
    SRR5139362'...
2025-09-18T17:20:42 prefetch.3.2.1:  SRA Normalized
    Format file is being retrieved
2025-09-18T17:20:42 prefetch.3.2.1:  Downloading via
    HTTPS...
2025-09-18T17:22:00 prefetch.3.2.1:  HTTPS download
    succeed
2025-09-18T17:22:00 prefetch.3.2.1:  'SRR5139362' is
    valid: 436693323 bytes were streamed from 436683321
2025-09-18T17:22:00 prefetch.3.2.1: 1) 'SRR5139362' was
    downloaded successfully
2025-09-18T17:22:00 prefetch.3.2.1: 1) Resolving '
    SRX2458127's dependencies...
2025-09-18T17:22:00 prefetch.3.2.1: 'SRX2458127' has 0
    unresolved dependencies
```

Formatando arquivo de sequenciamento para obtermos sua qualidade:

```
fastq-dump SRR5139362/SRR5139362.sra
```

```
Read 3430059 spots for SRR5139362/SRR5139362.sra
Written 3430059 spots for SRR5139362/SRR5139362.sra
```

Exibindo a saída:

```
head SRR5139362.fastq
```

```
@SRR5139362.1 1 length=200
CTCACGTCGCTTCTGACGTTTCGTTCCTGATATTTGCATACCCGTGACGTTTCT
ACTGACTTATGTGTTTATTGTACTGATCTCCAGACCTCCTTTTTTAAAAAAAGG
AGGTCTGGAGATCAGTACAATAAACACATAAGTCAGTAGAAACGTCACGGGTAT
```

```
GCAAATATCAGGAACGAAACGTCAGAAGCGACGTGAGA

+SRR5139362.1 1 length=200
B@@F=DDDHHHHHJIJJJHJJJJHIJJJJJJJJJJJJJJJJJJHIJJJFHIIJ
IJJIIJJHFGHEECDEFFFEDADEEEEDEDDCDDDBCDDDDDCDBBCCCFFFFF
HHHHHJIJIJJJJJJIIJJJJJJJJJJJJJJIJJJIIIJJJHIJJIJJIJ:BH
HHFFFFFFFEEEDEDDDDDDDABDDDDDDBBDDBDDDD

@SRR5139362.2 2 length=199
TTGAATCCTGTATTTAAAGAATTGGATTTCCAATAATAAATTTCTTTTGGTGAT
CTGTTTAGTGAAAAATGTCATTGACTTTACACCCTCTTCCGCATTCTGGAAGGG
CTATGTTTAAAATATTAATGCGGAAGAGGGTGTAAAGTCAATGACATTTTTCAC
TAAACAGATCACCAAAAGAAATTTATTATTGGAAATC

+SRR5139362.2 2 length=199
1=D4ADDHHHHHJJJJJJIJJIJJJIJIJJJJJJJJJJIIJJJJJIIJJJBGHI
IIJFGIIJGIIIJJIJJIIIJJEHIIGIIGGFHCDEDDEEDDBBBBCCCFFFFFH
HHHHJIJJJJJJJJJJJJJJJJJHJJIJJJBFFHJJJIIJIHJJJJJJIJJJIJ
JJJHHHHGFFFFFECDDEDDDDDDDEDDFEDCDDDDD

@SRR5139362.3 3 length=199
CTTTTCCTTCAATTCGTTTATTATATTTAATAGAAGTATTTGAACAGCCAGAAT
TATCAATTAAAGTTATAGGTCATCAATGATATTGATCTTACGAGTATTGATTTA
AACGTCCTGGAATTGCATCTACTTTTACTCCTAAAGAAGGAATAGTTCAAGAAT
GAATTACATCTGATCTTGAAACAATTATGCGAATATC
```

## 2.2 Análise de Qualidade Inicial

Utilizou-se o programa FASTQC para avaliar a qualidade das reads no arquivo FASTQ original.

## 2.3 Processamento das Reads

Foram aplicados diferentes parâmetros de corte utilizando o EA-Utils (fastq-mcf).

**Corte 1: Filtro de Qualidade (Q35)**

```
fastq-mcf -q 35 n/a SRR5139362.fastq -o SRR5139362_Q35.
   fastq
```

**Corte 2: Filtro de Comprimento (l30 + L150)**

```
fastq-mcf -l 30 -L 150 n/a SRR5139362.fastq -o
   SRR5139362_l30_L150.fastq
```

**Corte 3: Filtros Combinados (Q35 + l30 + L150)**

```
fastq-mcf -q 35 -l 30 -L 150 n/a SRR5139362.fastq -o
   SRR5139362_Q35.fastq
```

# 3 Resultados e Análise

## 3.1 Qualidade Inicial das Reads

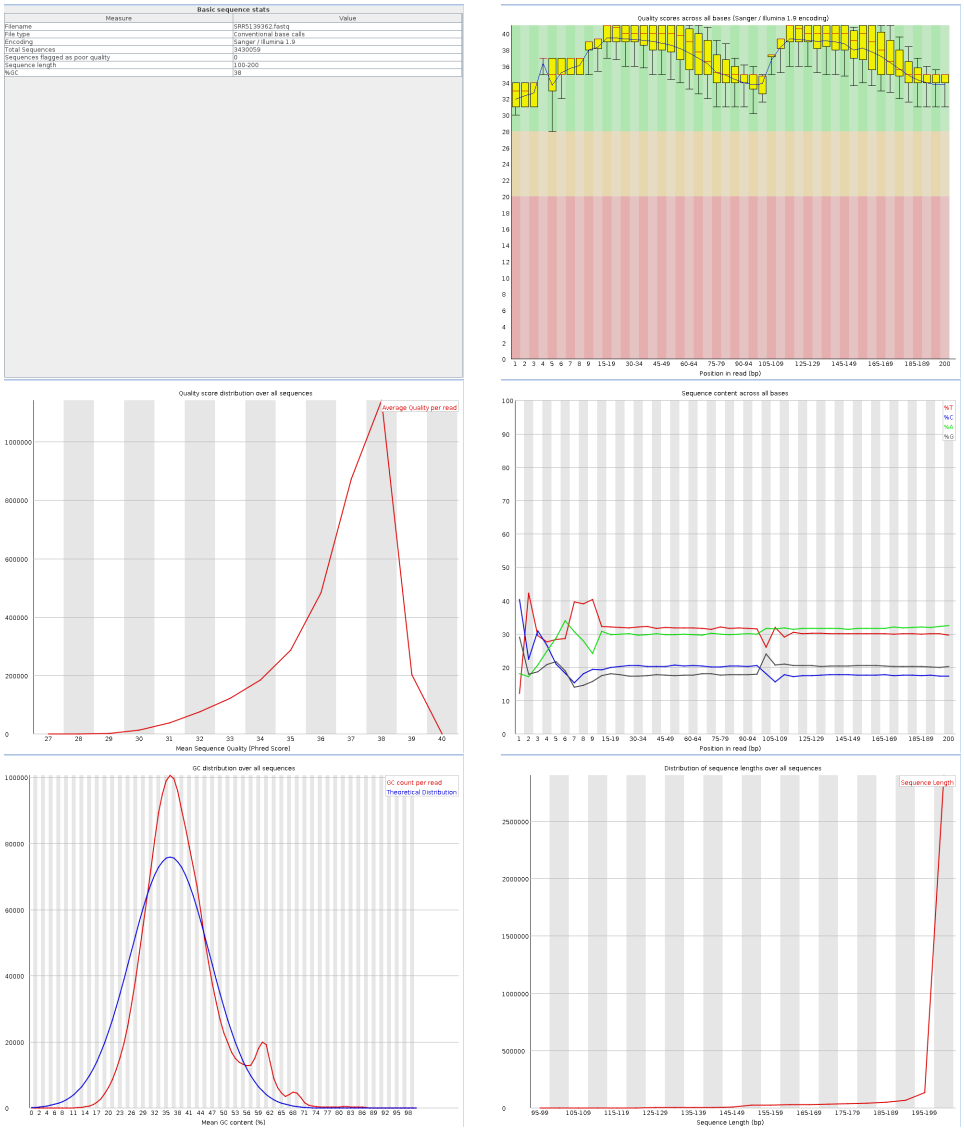


Figura 2: Qualidade da read inicial

## 3.2 Análise Comparativa dos Processamentos

### 3.2.1 Impacto do Filtro de Qualidade (Q35)

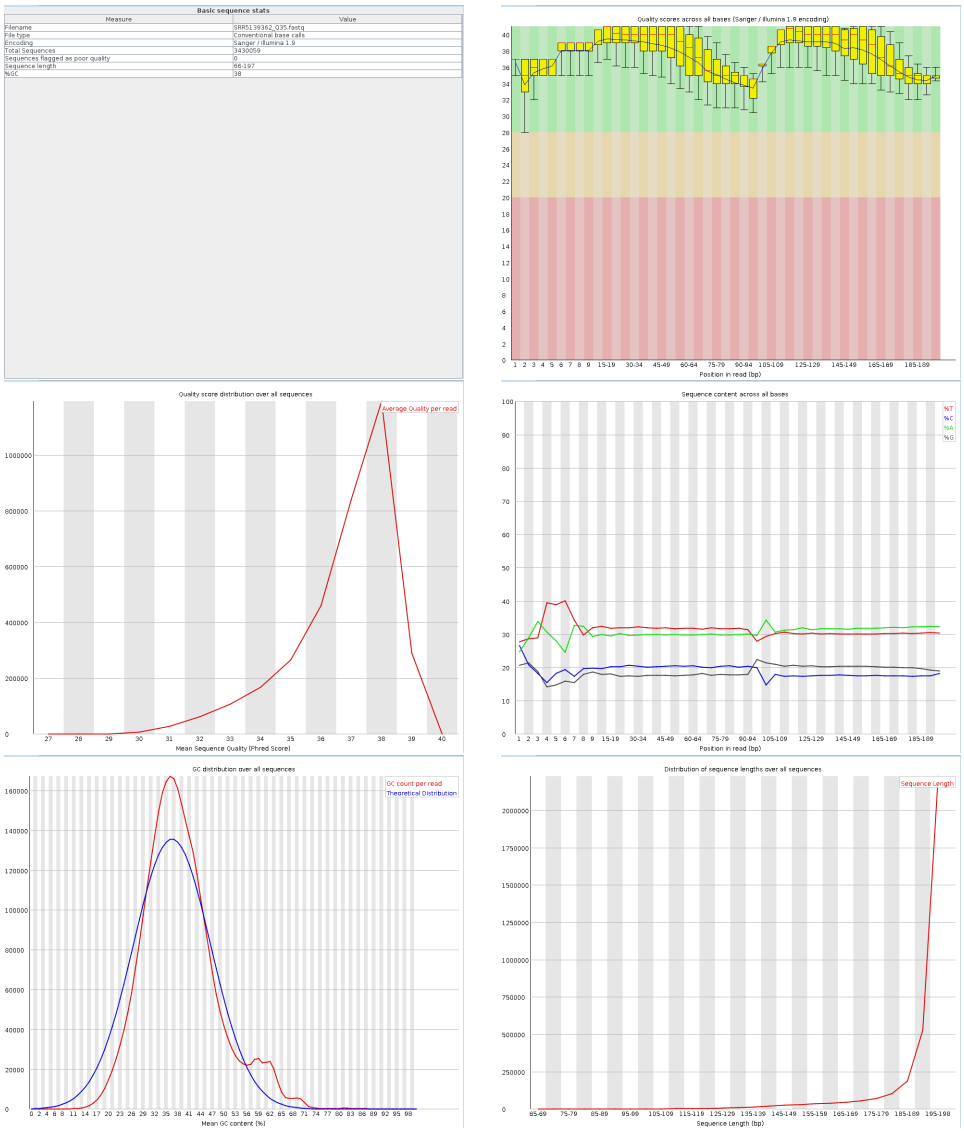


Figura 3: Qualidade da read Q35

### 3.2.2 Impacto do Filtro de Comprimento (l30 + L150)

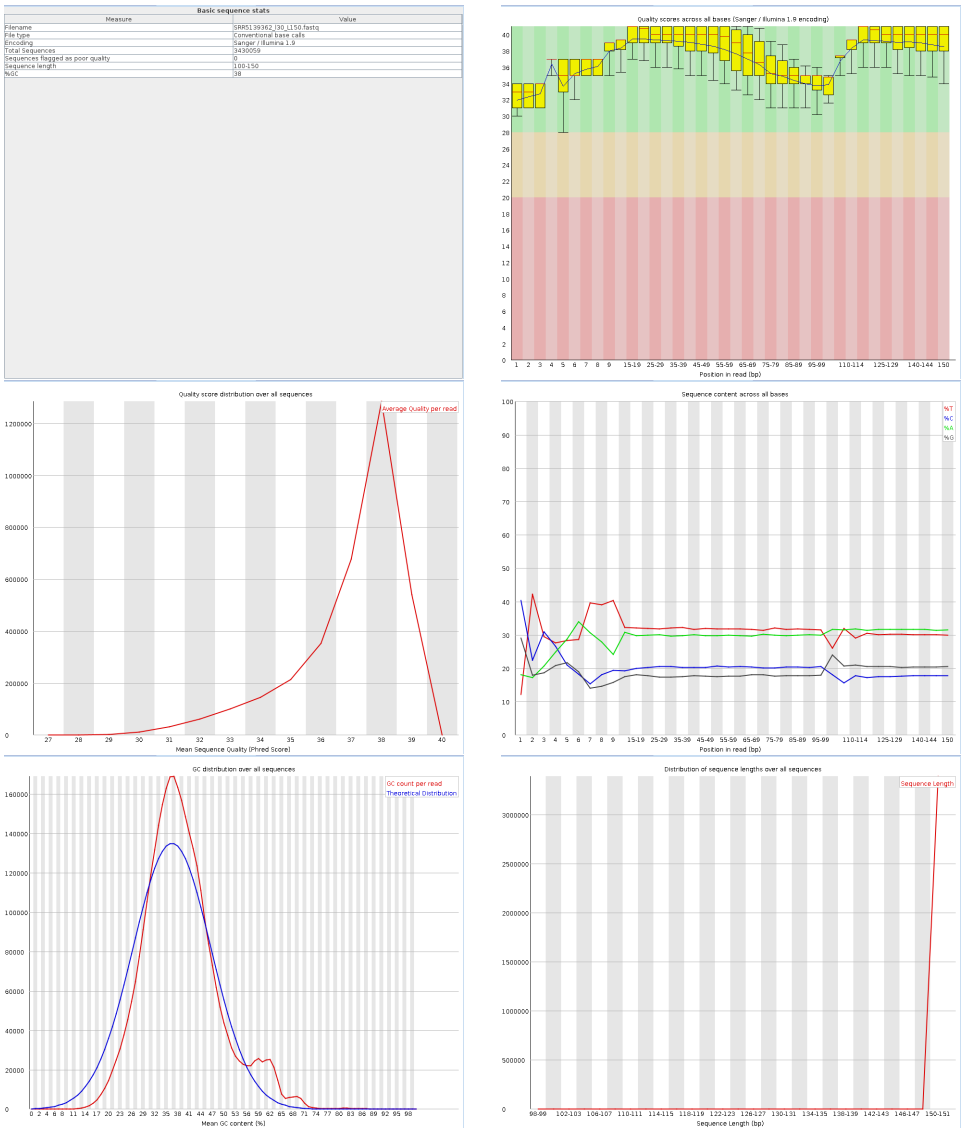


Figura 4: Qualidade da read l30 + L150

### 3.2.3 Impacto dos Filtros Combinados (Q35 + l30 + L150)

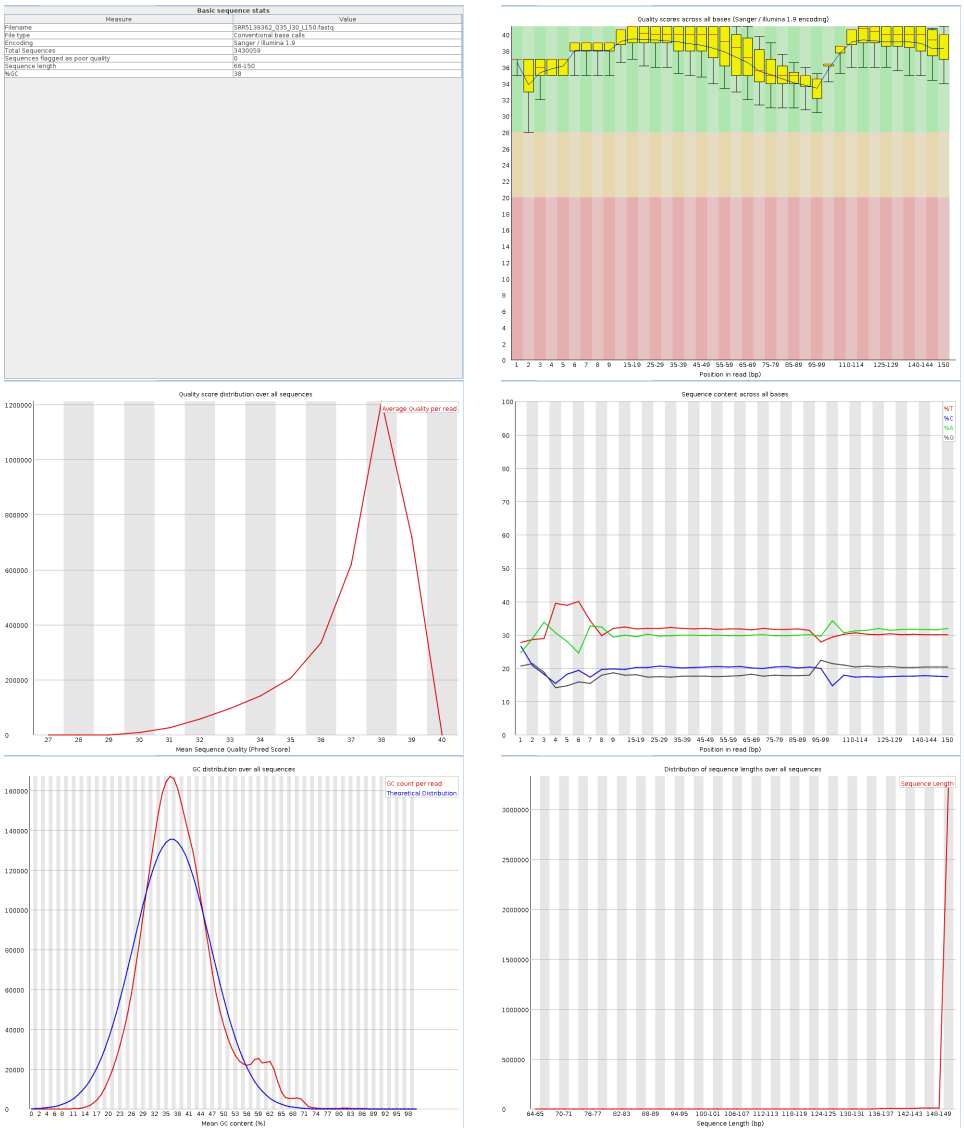


Figura 5: Qualidade da read Q35 + l30 + L150

# 4 Resultados e Análise

## 4.1 Interpretação dos Quality Scores

Conforme a escala Phred (fastsanger), um quality score Q35 corresponde a uma probabilidade de erro de aproximadamente 0,03%, ou uma precisão de 99,97%. Isso significa que escolhemos manter apenas bases com altíssima confiabilidade.

## 4.2 Critérios de Corte Aplicados

**Q35**: Embora o material já apresentasse qualidade superior a Q30 (considerada boa), optou-se por Q35 para garantir máxima precisão na análise downstream, especialmente importante para estudos de expressão gênica onde variações pequenas podem ser biologicamente relevantes.
**30 + L150**: A aplicação de filtros de comprimento em ambas as extremidades visa otimizar a qualidade do alinhamento. Reads muito curtas têm maior probabilidade de alinhar em múltiplas posições no genoma por acaso, gerando mapeamentos ambíguos. Por outro lado, reads excessivamente longas podem conter mais erros sequenciais acumulados e podem apresentar problemas de qualidade nas extremidades. A janela de 30-150 nucleotídeos procura garantir especificidade de alinhamento sem comprometer a qualidade geral dos dados.
**Desvantagens**: Potencial perda de informação biológica relevante (reads descartadas), possível redução na cobertura de transcritos expressos em baixos níveis

# 5 Conclusão

A análise de qualidade das reads de RNA-Seq de *Nephila clavipes* (SRX2458127) demonstrou que os filtros aplicados foram eficazes na otimização dos dados para análises downstream. O dataset original já apresentava alta qualidade ($Q > 30$), confirmando a robustez dos dados utilizados no estudo de Babb et al. (2017) sobre genes de seda desta espécie.
Os resultados da aplicação dos filtros combinados (Q35 + l30 + L150) revelaram melhorias específicas e importantes:
**Melhoria na qualidade por sequência**: O aumento no score de qualidade por sequência de 1.000.000 para 1.200.000 demonstra que os filtros

foram seletivos, mantendo preferencialmente as reads de maior confiabilidade individual.
**Otimização do conteúdo GC**: O incremento no conteúdo GC por sequência indica que foram removidas reads com composição nucleotídica atípica, preservando sequências mais representativas do transcriptoma real de N. clavipes.
**Padronização do comprimento**: A redução do comprimento predominante de 195-199 para 148-149 nucleotídeos reflete diretamente o efeito do filtro L150, resultando em maior uniformidade das reads e adequação para alinhamento específico.
**Estabilidade dos demais parâmetros**: A manutenção inalterada dos outros indicadores confirma que a estratégia de filtragem foi precisa, não introduzindo artefatos ou vieses indesejados.
Para o contexto específico desta pesquisa - caracterização da diversidade de genes de seda - a qualidade final obtida é adequada para análises de expressão diferencial e identificação de transcritos relacionados à produção dos diferentes tipos de seda.

# 6 Referências

# Referências


[1] Babb, P., Lahens, N., Correa-Garhwal, S. et al. The *Nephila clavipes* genome highlights the diversity of spider silk genes and their complex expression. *Nat Genet* **49**, 895–903 (2017). `https://doi.org/10.1038/ng.3852`.

[2] UNIVERSIDADE DE SÃO PAULO. Disciplina: Introdução à Bioinformática - SCC0271. Material de aula: Qualidade do Sequenciamentos. Plataforma e-disciplinas USP, 2025. Disponível em: `https://edisciplinas.usp.br`. Acesso em: 18/09/2025.