Treating **E**volutionarily heterogeneous **C** type **H**epatitis

Genetic Heterogeneity of Hepatitis C Virus in Association with Antiviral Therapy Determined by Ultra-Deep Sequencing
Akihiro Nasu et al.

Bioinformatics Bootcamp
September 24, 2013

Stephanie Feupe
Shamim Mollah
Jenhan Tao
Mengchi Wang

Overview

1. Hepatitis C Background

2. Paper Overview
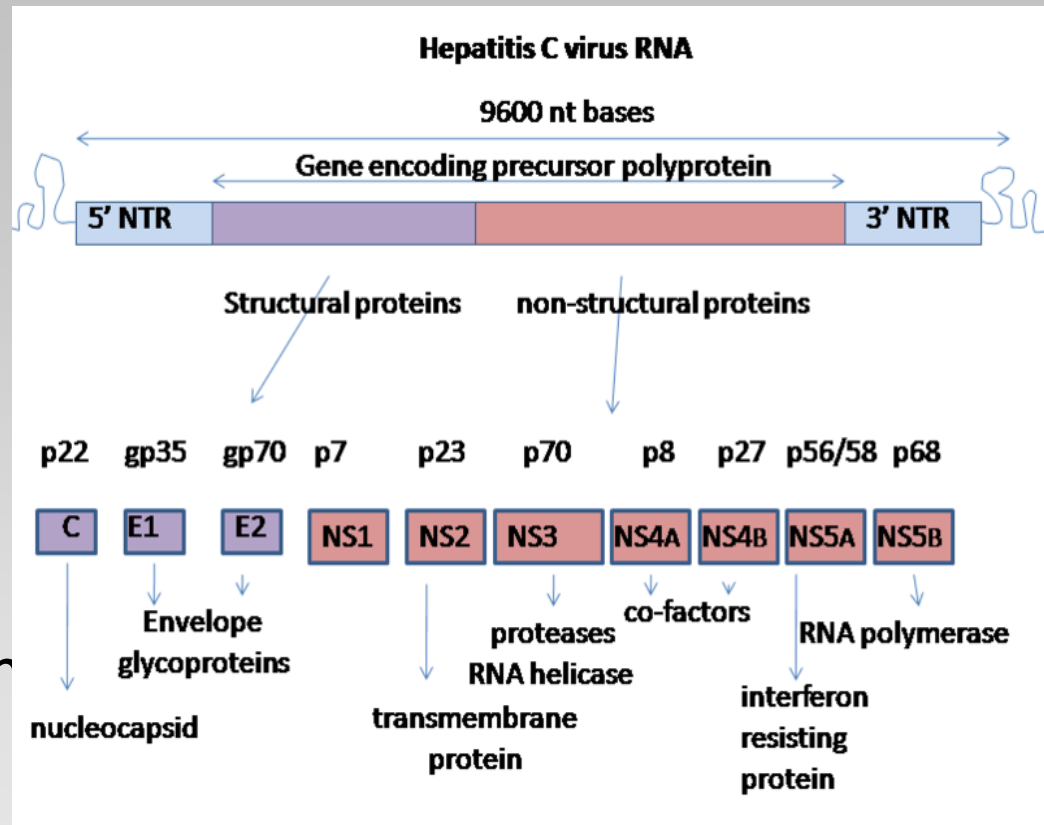
3. Methodology

4. Results

5. Conclusion

Hepatitis C

- Hepatitis C is an infectious disease, affecting primarily the liver

  - Can lead to scarring of the liver and cirrhosis if not treated

- Passed through contact with contaminated blood and bodily fluids

- Often asymptomatic

Hepatitis C Virus

- An enveloped RNA virus of approximately 9.6-kb single stranded RNA genome

- Highly mutable and can escape immune detection

- 11 HCV genotypes with several distinct subtypes have been identified throughout the world (WHO)

Hepatitis C virus RNA

9600 nt bases

Gene encoding precursor polyprotein

| 5' NTR | | | 3' NTR |

Structural proteins   non-structural proteins

p22   gp35   gp70   p7   p23   p70   p8   p27   p56/58   p68

| C | E1 | E2 | NS1 | NS2 | NS3 | NS4A | NS4B | NS5A | NS5B |

Envelope glycoproteins

co-factors

proteases

RNA polymerase

RNA helicase

nucleocapsid

transmembrane protein

interferon resisting protein

Treatment

- HCV genotype 1 is treated with three drugs: **pegylated interferon**, **ribavirin**, and a **protease inhibitor**.

- HCV genotypes 2 and 3 are treated with **pegylated interferon** and **ribavirin** (standard dose is 800 mg/day, weight based dosing).

- HCV genotype 4 is treated with **pegylated interferon** and weight-based **ribavirin**;

- Different genotypes vary in their responsiveness to interferon/ ribavirin combination therapy

- Treatment isn't always necessary

Genetic Heterogeneity of Hepatitis C Virus in Association with Antiviral Therapy Determined by Ultra-Deep Sequencing

Heterogeneity vs. Treatment

- **Heterogeneity**: genomic diversity among the virus.

- 27 patients are treated with interferon therapy.

- Deep sequencing before and after the therapy.

- The paper analysis how the treatment decreased heterogeneity.
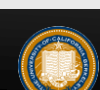
Dataset

Deep sequencing:

- Isolated HCV quasi-species from patients' serum

- Fragmented for deep sequencing

- A representative HCV sequence is determined as reference

- Note:

  - Start with .sra files

  - No human genome. Define our reference sequence.

Project Goal

- How the treatment impact the **overall** genome heterogeneity?

- What about other unidentified regions of HCV? (Unbiased alignment)

- What do these region correspond to? (NCBI Blast)

- Potential drug design based on stable regions in HCV genome?

Project Strategy

- Genome reference
  - Hepatitis C virus gene for polyprotein, genomic RNA, isolate Equatorial Guinea (9616 base pair)

- Strategy:
  1. Select 10 patients randomly
  2. For pre/post treatment, produce alignment
  3. Independent two-sample T-Test
  4. Perform BED analysis
  5. Parse alignment for sequences of interest
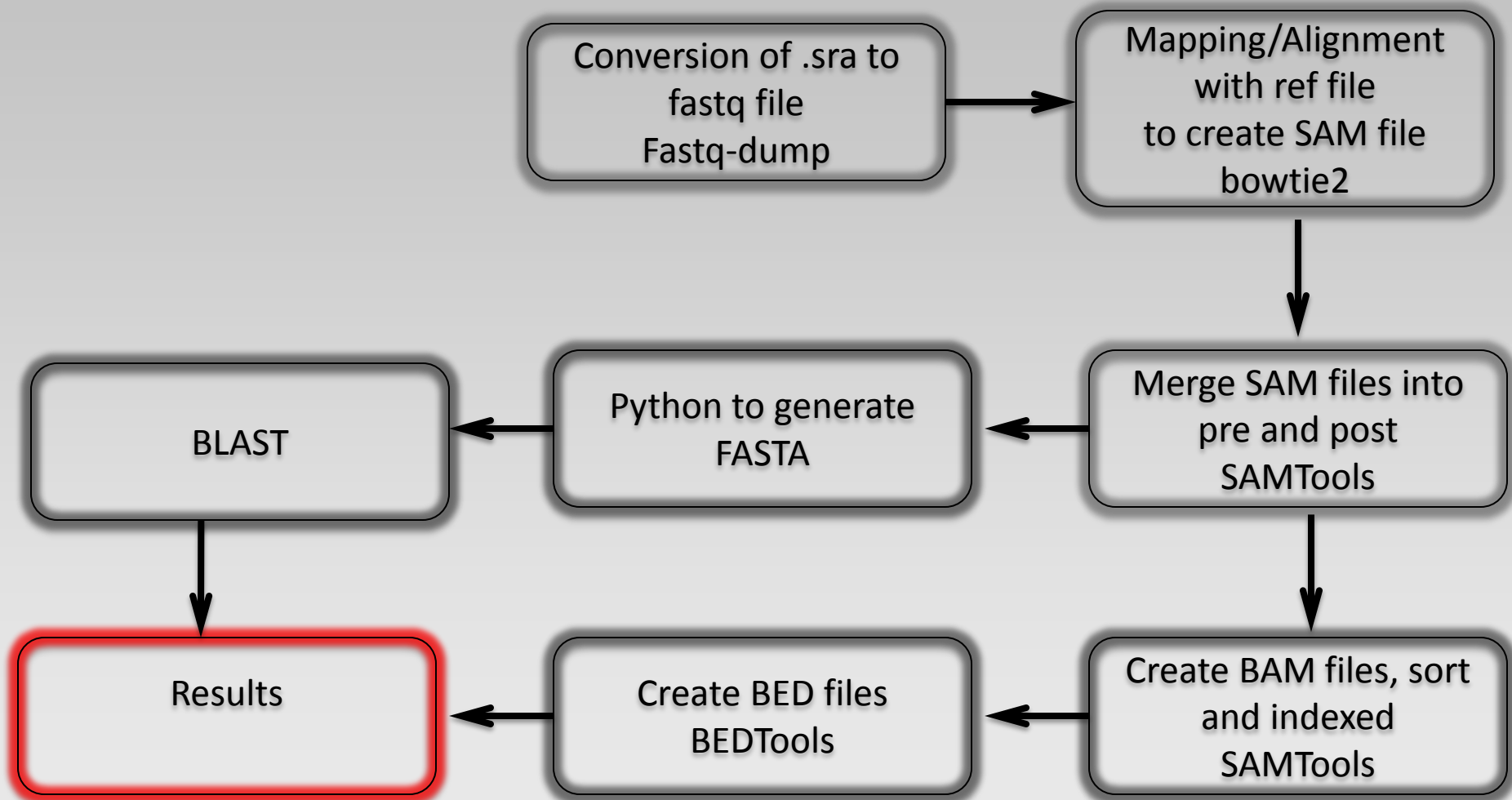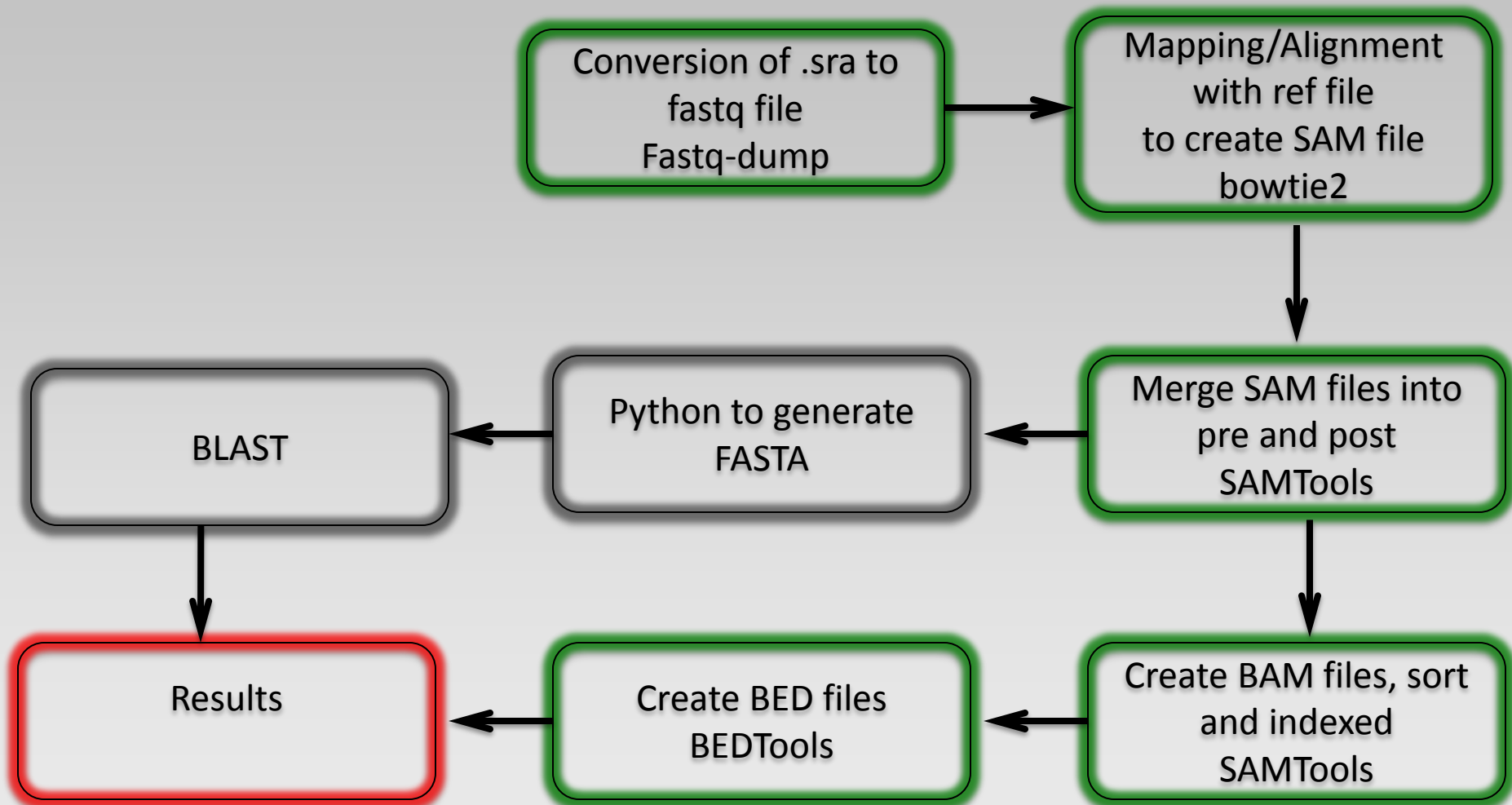 6. BLAST sequences of interest

Method

- Which regions of the virus was conserved
- Input data files
  - 10 pre and post treatment data files (ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/DRR)
- Genome reference
  - Hepatitis C virus gene for polyprotein, genomic RNA, isolate Equatorial Guinea (9616 nucleotides)
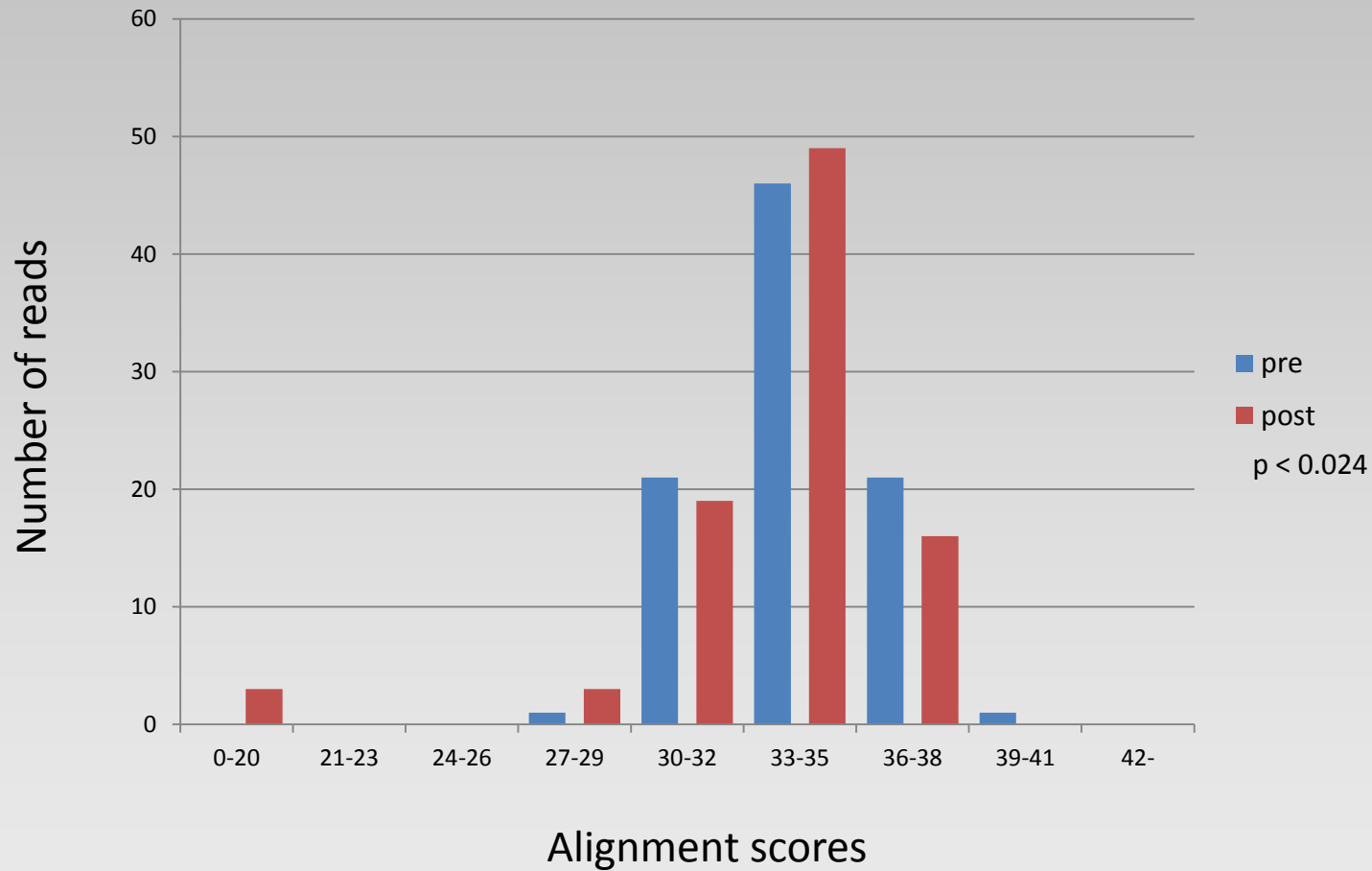- Independent two-sample T-Test
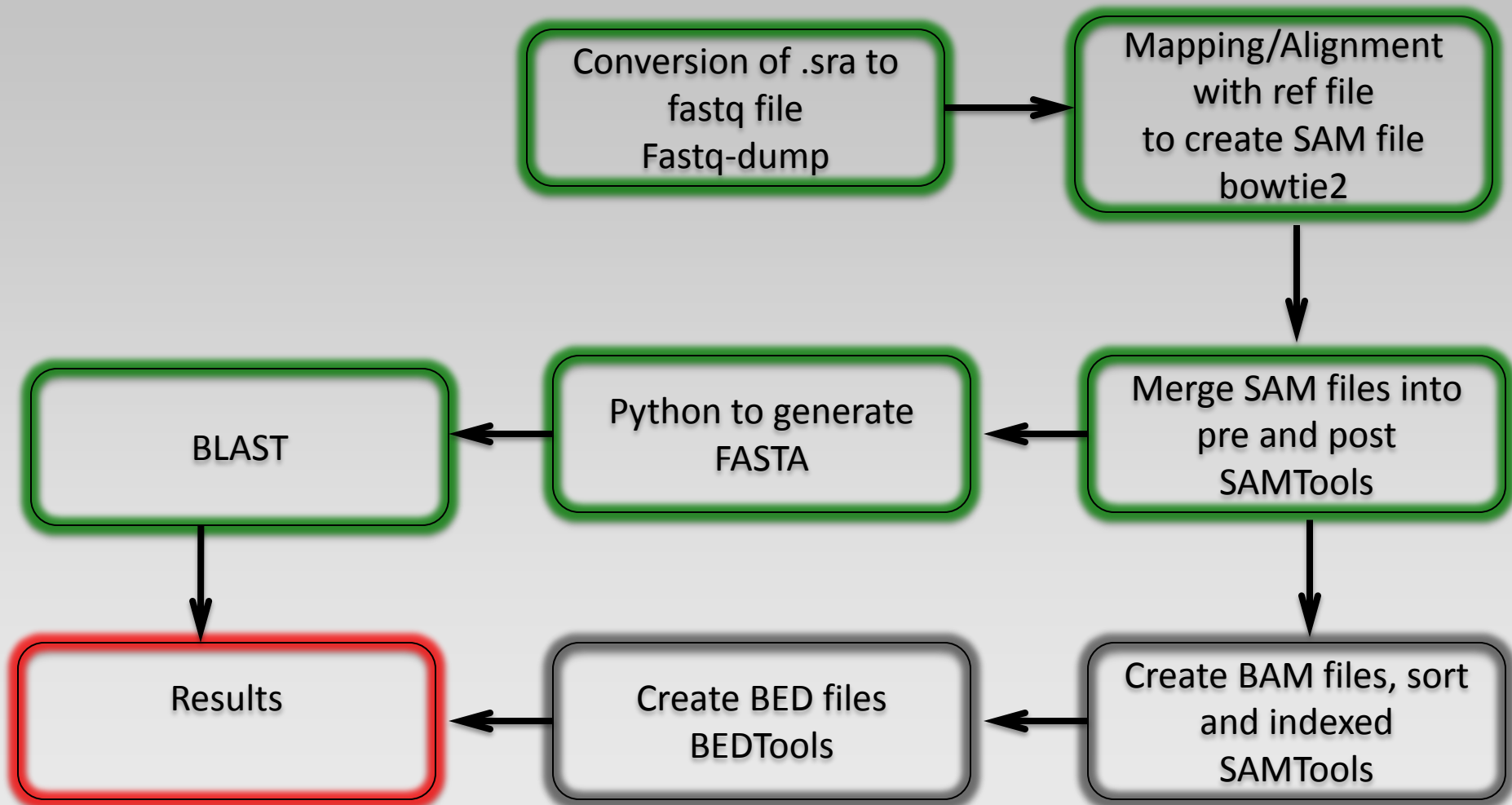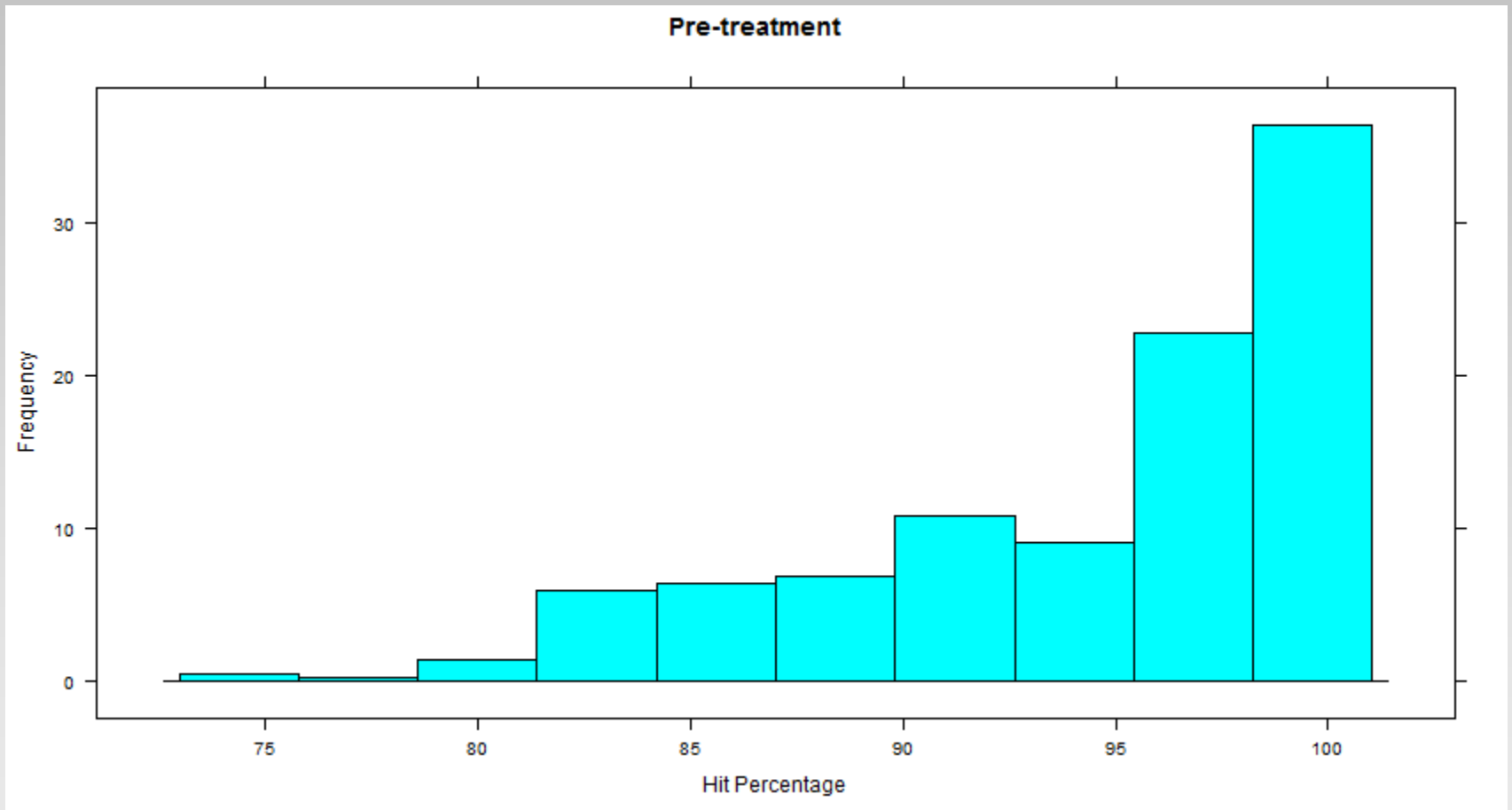- BLAST

Analysis Pipeline

BED Analysis

```
┌─────────────────────┐         ┌─────────────────────┐
│  Conversion of .sra │────────▶│  Mapping/Alignment  │
│     to fastq file   │         │     with ref file   │
│     Fastq-dump      │         │  to create SAM file │
└─────────────────────┘         │       bowtie2       │
                                └─────────────────────┘
                                           │
                                           ▼
┌───────────┐   ┌─────────────────┐   ┌─────────────────────┐
│   BLAST   │◀──│ Python to       │◀──│  Merge SAM files    │
│           │   │ generate FASTA  │   │  into pre and post  │
└───────────┘   └─────────────────┘   │      SAMTools       │
      │                               └─────────────────────┘
      ▼                                          │
┌───────────┐   ┌─────────────────┐   ┌─────────────────────┐
│  Results  │◀──│ Create BED files│◀──│ Create BAM files,   │
│           │   │    BEDTools     │   │  sort and indexed   │
└───────────┘   └─────────────────┘   │      SAMTools       │
                                      └─────────────────────┘
```

BED Results

Blast Analysis

Conversion of .sra to fastq file
Fastq-dump
→
Mapping/Alignment with ref file
to create SAM file
bowtie2
↓
Merge SAM files into pre and post
SAMTools
←
Python to generate FASTA
←
BLAST
↓
Results
←
Create BED files
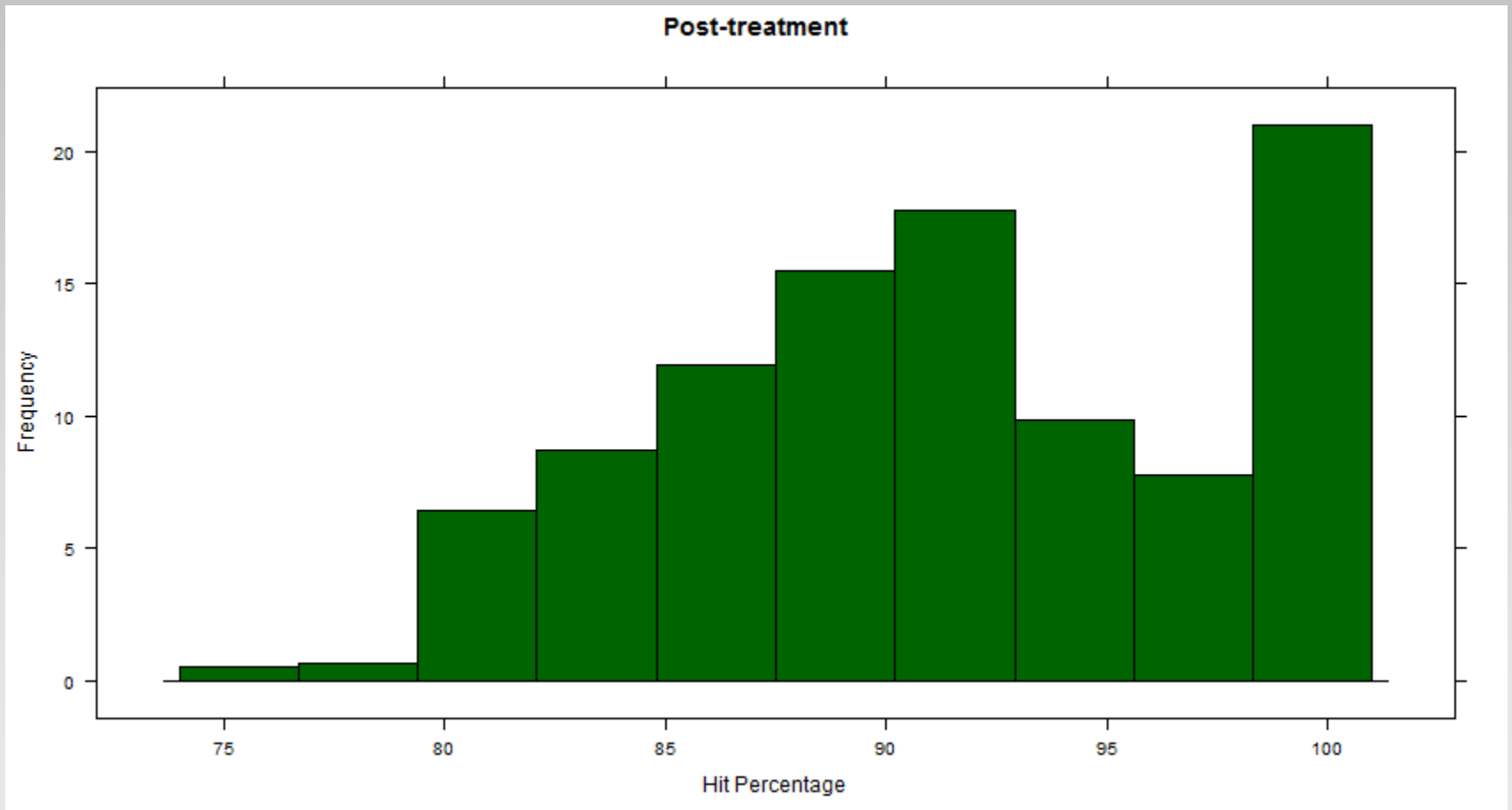BEDTools
←
Create BAM files, sort and indexed
SAMTools

Blast Results

Blast Results
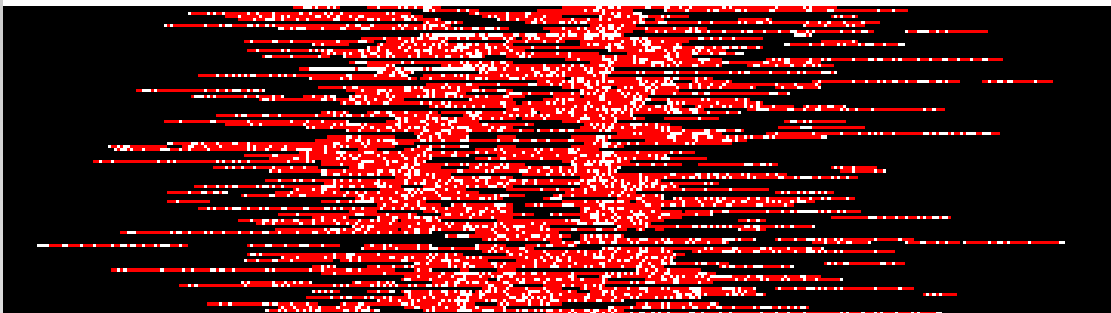

Post-treatment

Blast Results

- Hepatitis C virus subtype 1b NS5B gene for polyprotein
- Hepatitis C virus genomic RNA
- Hepatitis C virus subtype 1b strain MD7-1
- Hepatitis C virus NS5b gene
- Hepatitis C virus subtype 1b strain HeBei
- Hepatitis C virus subtype 1b strain HCV-L2
- Hepatitis C virus isolate V3112 NS3-4 protease gene

Mapping to HCV gene database
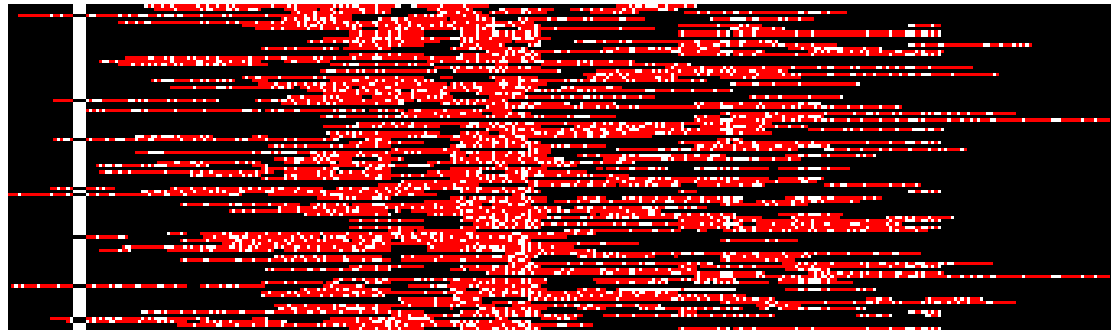
Pre-treatment

Post-treatment

Legend

□ Matches

▬ Mismatches

▬ Indels

Conclusion

- Identified many conserved regions despite reduced population size

- Treatments that target conserved regions may be more affective to a greater variety of HCV strains

- Todo:
  - Automated processing of BLAST hits
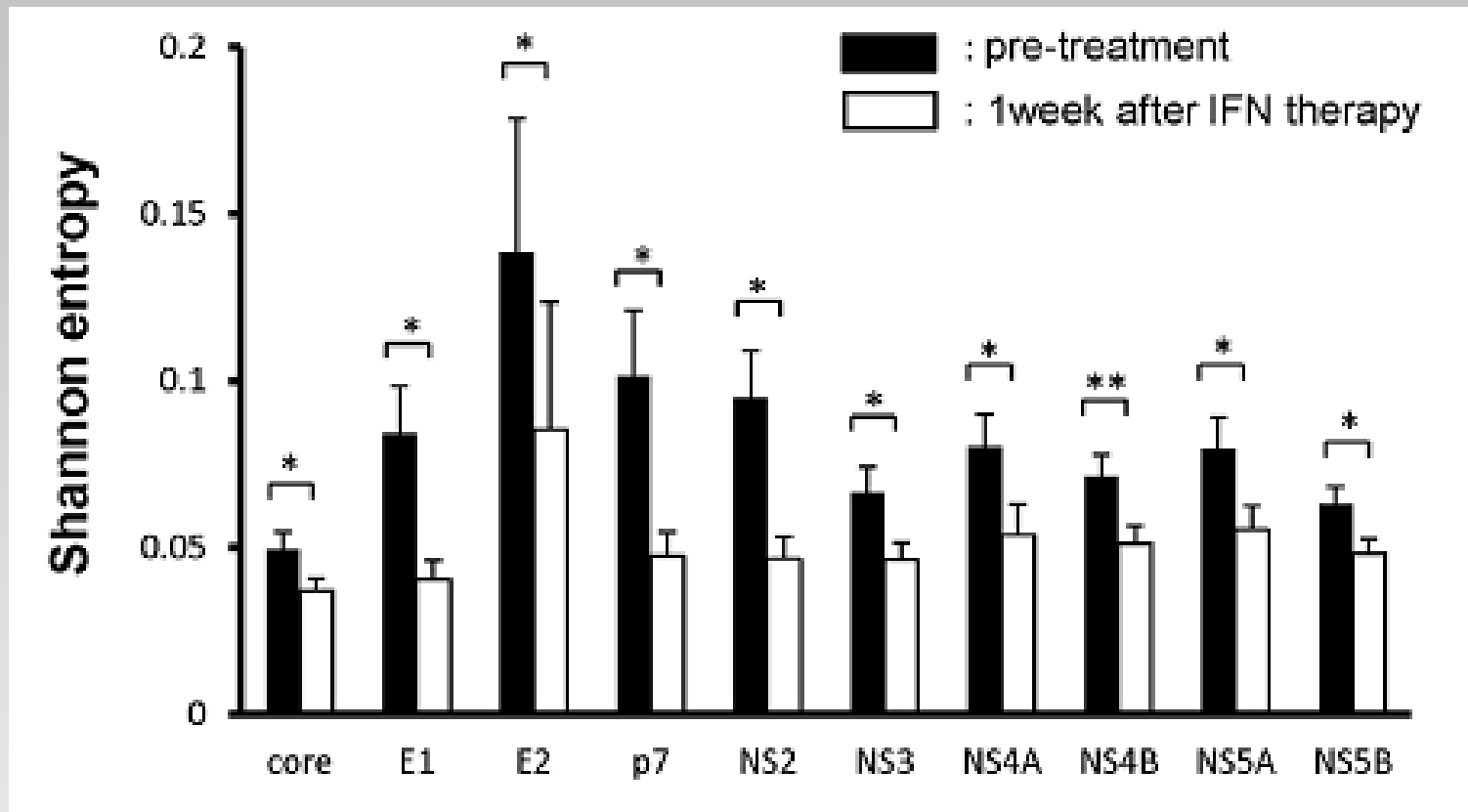 - Optimize pipeline for scalability

Conclusion

- Todo:
  - Automated processing of BLAST hits
  - Optimize pipeline for scalability

Questions?

Results

-Heterogeneity is decreased after treatment.
-However…