

# Science: An Explosion Of Bioinformatics Careers

## Programming

Additional critical skills are required for big data careers in industry, such as text mining, ontology, data integration, machine learning, and information architecture. A superior “quantitative ability,” as Gentleman calls it, which covers a range of statistical capabilities, is a must, as are overarching computing skills. These include core programming abilities, such as coding in C++ or Java, or scripting in PERL or Python, says Van Criekinge. It is vital to be able to navigate operating systems like UNIX and Linux as well as have knowledge of common tools such as Hadoop and NoSQL databases, adds Mohan. Experience in data visualization and building effective user interfaces, as well as familiarity with hardware, buttresses your marketability.

## Assignment of Perl

- Write a program to print Fibonacci sequence. The length of output sequence is specified by the first command line parameter. (fibonacci.pl)
- Try to optimize the pos\_annotate.pl as much as you can. (pos\_annotateV3.pl)

# Perl in Bioinformatics

Gang Chen  
chengang@bgitecholutions.com

October 11, 2014

# Outline

- 1 Perl Modules
- 2 SNP Annotation
- 3 Bioinformatics Project in Perl

# Next

## 1 Perl Modules

- Usage of Perl Modules
- Management of Perl Modules
- Development of Perl Modules
- Submit your modules to CPAN

## 2 SNP Annotation

## 3 Bioinformatics Project in Perl

# Next

## 1 Perl Modules

- Usage of Perl Modules
- Management of Perl Modules
- Development of Perl Modules
- Submit your modules to CPAN

## 2 SNP Annotation

- A Perl Script for SNP Annotation
- SNPAnno module

## 3 Bioinformatics Project in Perl

- Annovar
- Cirocs
- BioPerl

# Web Scraping in Perl

## LWP

The libwww-perl collection is a set of Perl modules which provides a simple and consistent application programming interface (API) to the World-Wide Web.  
<https://metacpan.org/pod/LWP>

# Example: lwp.pl

## Files

- pubmedids.txt: A list of pubmed ids
- lwp.pl: Get publication titles of these ids from PubMed.



# Next

## 1 Perl Modules

- Usage of Perl Modules
- **Management of Perl Modules**
- Development of Perl Modules
- Submit your modules to CPAN

## 2 SNP Annotation

- A Perl Script for SNP Annotation
- SNPAnno module

## 3 Bioinformatics Project in Perl

- Annovar
- Cirocs
- BioPerl

# Searching

[www.CPAN.org](http://www.CPAN.org)

# Installation

- cpan command: Linux, Mac OS and Windows(Strawberry)
- perl -MCPAN -e shell
- cpanm
- from source
- ppm: ActivePerl

# Next

## 1 Perl Modules

- Usage of Perl Modules
- Management of Perl Modules
- **Development of Perl Modules**
- Submit your modules to CPAN

## 2 SNP Annotation

- A Perl Script for SNP Annotation
- SNPAnno module

## 3 Bioinformatics Project in Perl

- Annovar
- Cirocs
- BioPerl

# Development

see PUBMED.pm and run.pl

# Next

## 1 Perl Modules

- Usage of Perl Modules
- Management of Perl Modules
- Development of Perl Modules
- **Submit your modules to CPAN**

## 2 SNP Annotation

- A Perl Script for SNP Annotation
- SNPAnno module

## 3 Bioinformatics Project in Perl

- Annovar
- Cirocs
- BioPerl

# Why?

- Easy the installation of your module
- Share your work with the community
- Increase your importance in the community
- Get response form the community to improve your module

# How?

<http://www.cpan.org/modules/04pause.html>



# Next

1

Perl Modules

2

**SNP Annotation**

- A Perl Script for SNP Annotation
- SNPAnno module

3

Bioinformatics Project in Perl

# SNP Annotation

## Task

Given a list of genome positions, add corresponding gene symbol to each position.

## Files

- Example input file: pos.txt
- Annotation database: refGene.txt
- Example script: pos\_annotate.pl
- Optimized Script: pos\_annotateV2.pl

# Next

## 1 Perl Modules

- Usage of Perl Modules
- Management of Perl Modules
- Development of Perl Modules
- Submit your modules to CPAN

## 2 SNP Annotation

- A Perl Script for SNP Annotation
- SNPAnno module

## 3 Bioinformatics Project in Perl

- Annovar
- Cirocs
- BioPerl

# pos\_annotate.pl

# pos\_annotateV2.pl

# Next

## 1 Perl Modules

- Usage of Perl Modules
- Management of Perl Modules
- Development of Perl Modules
- Submit your modules to CPAN

## 2 SNP Annotation

- A Perl Script for SNP Annotation
- SNPAnno module

## 3 Bioinformatics Project in Perl

- Annovar
- Cirocs
- BioPerl

# SNPAnno module

- SNPAnno.pm
- run\_SNPAnno.pl

# Next

- 1 Perl Modules
- 2 SNP Annotation
- 3 Bioinformatics Project in Perl**
  - Annovar
  - Ciros
  - BioPerl



# Next

## 1 Perl Modules

- Usage of Perl Modules
- Management of Perl Modules
- Development of Perl Modules
- Submit your modules to CPAN

## 2 SNP Annotation

- A Perl Script for SNP Annotation
- SNPAnno module

## 3 Bioinformatics Project in Perl

- **Annovar**
- Cirocs
- BioPerl

# Annovar

## Annovar

ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data

<http://www.openbioinformatics.org/annovar/>

# Next

## 1 Perl Modules

- Usage of Perl Modules
- Management of Perl Modules
- Development of Perl Modules
- Submit your modules to CPAN

## 2 SNP Annotation

- A Perl Script for SNP Annotation
- SNPAnno module

## 3 Bioinformatics Project in Perl

- Annovar
- **Cirocs**
- BioPerl

# Circos

## Circos

Circos is a software package for visualizing data and information. It visualizes data in a circular layout — this makes Circos ideal for exploring relationships between objects or positions.

<http://circos.ca>



# Next

## 1 Perl Modules

- Usage of Perl Modules
- Management of Perl Modules
- Development of Perl Modules
- Submit your modules to CPAN

## 2 SNP Annotation

- A Perl Script for SNP Annotation
- SNPAnno module

## 3 Bioinformatics Project in Perl

- Annovar
- Cirocs
- **BioPerl**

# BioPerl

<http://bioperl.org>

# Installation

- from source
- cpan (recommended)



# Thanks!