University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Architecture of large projects in bioinformatics (ADP)

Version control systems

Łukasz P. Kozłowski

lukaskoz@mimuw.edu.pl

Warsaw, 2025

1. Data formats in bioinformatics,
2. Popular software libraries (BioPerl, BioPython)
3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)
4. Software licensing for scientific purposes. Free-software licensing. Patents.

5. Generic model Organism database (GMOD) project - assumptions, history and usage

6. Genome browsers, problem description and state of the solutions

7. Version control systems (CVS, SVN, git), and online collaboration ad distribution platforms (github, sourceforge).

8. Software testing, automated testing frameworks.

9. Scientific workflow systems - taverna and galaxy. MyExperiment platform. Reproducible research.

10. Literate programming idea and sweave, markdown, software documentation.

11. Interactive scripting platforms, Rstudio, Jupyter.

1. Data formats in bioinformatics,
2. Popular software libraries (BioPerl, BioPython)
3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)
4. Software licensing for scientific purposes. Free-software licensing. Patents.

5. Generic model Organism database (GMOD) project - assumptions, history and usage

6. Genome browsers, problem description and state of the solutions

7. Version control systems (CVS, SVN, git), and online collaboration ad distribution platforms (github, sourceforge).

8. Software testing, automated testing frameworks.

9. Scientific workflow systems - taverna and galaxy. MyExperiment platform. Reproducible research.

10. Literate programming idea and sweave, markdown, software documentation.

11. Interactive scripting platforms, Rstudio, Jupyter.

1. Data formats in bioinformatics,
2. Popular software libraries (BioPerl, BioPython)
3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)
4. Software licensing for scientific purposes. Free-software licensing. Patents.

5. Generic model Organism database (GMOD) project - assumptions, history and usage

6. Genome browsers, problem description and state of the solutions

7. Version control systems (CVS, SVN, git), and online collaboration ad distribution platforms (github, sourceforge).

8. Software testing, automated testing frameworks.

9. Scientific workflow systems - taverna and galaxy. MyExperiment platform. Reproducible research.

10. Literate programming idea and sweave, markdown, software documentation.

11. Interactive scripting platforms, Rstudio, Jupyter.

Version control systems (CVS, SVN, git)

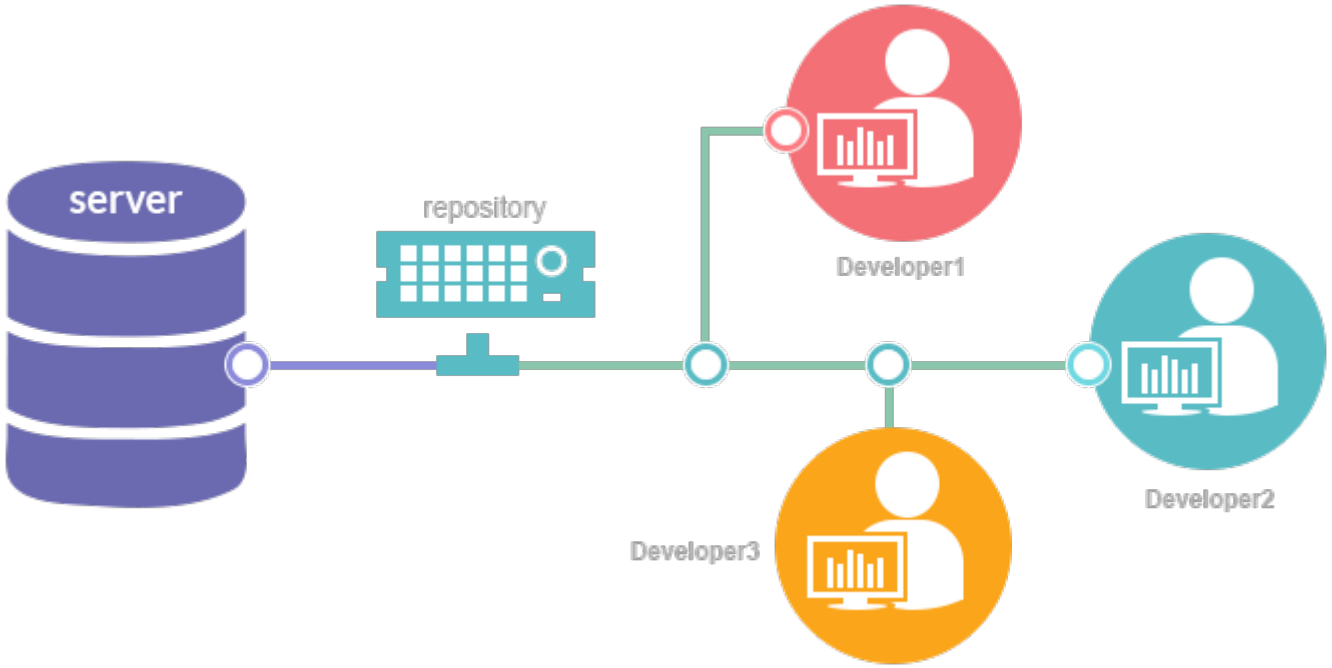
Online collaboration ad distribution platforms (github, sourceforge)

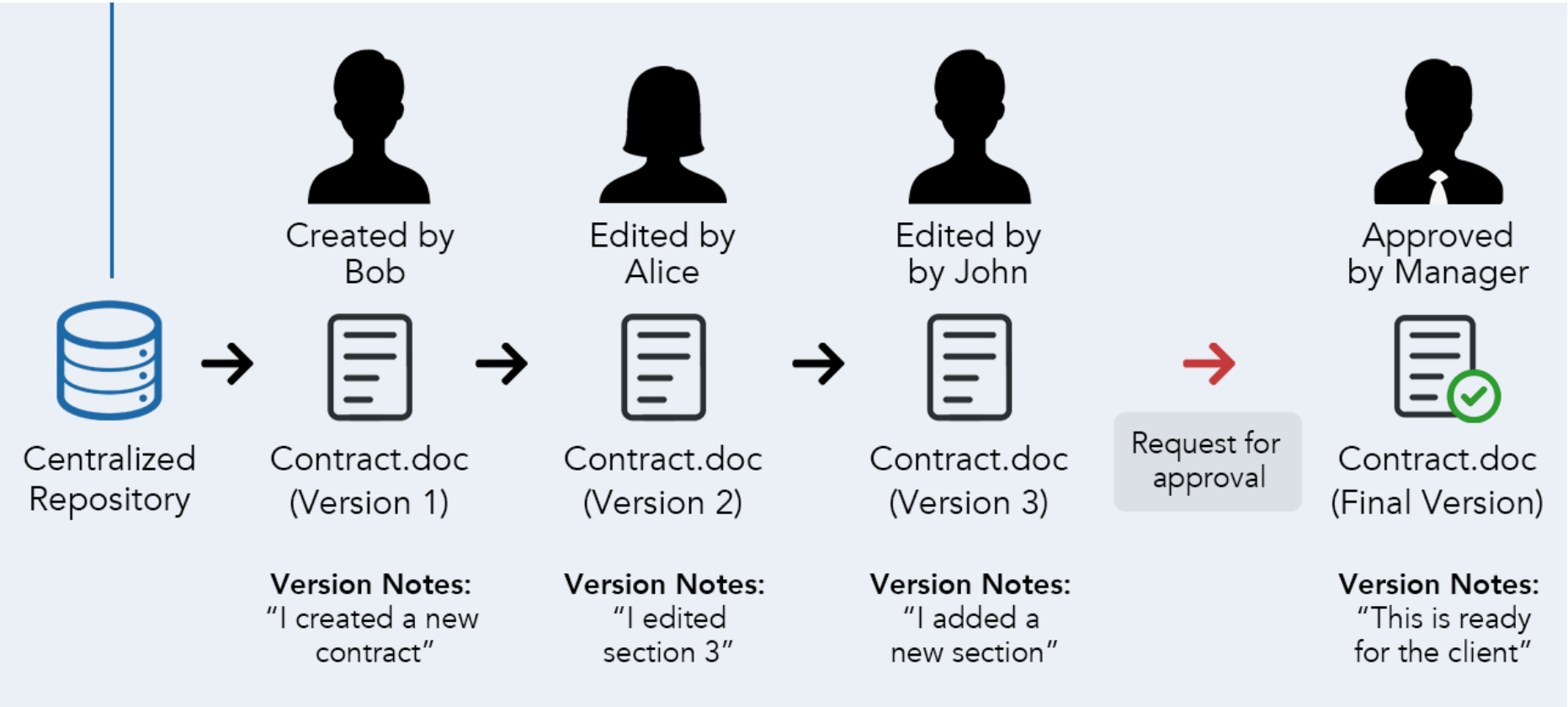
Version control systems (CVS, SVN, git)

Online collaboration ad distribution platforms (github, sourceforge)

What is a “version control system”?

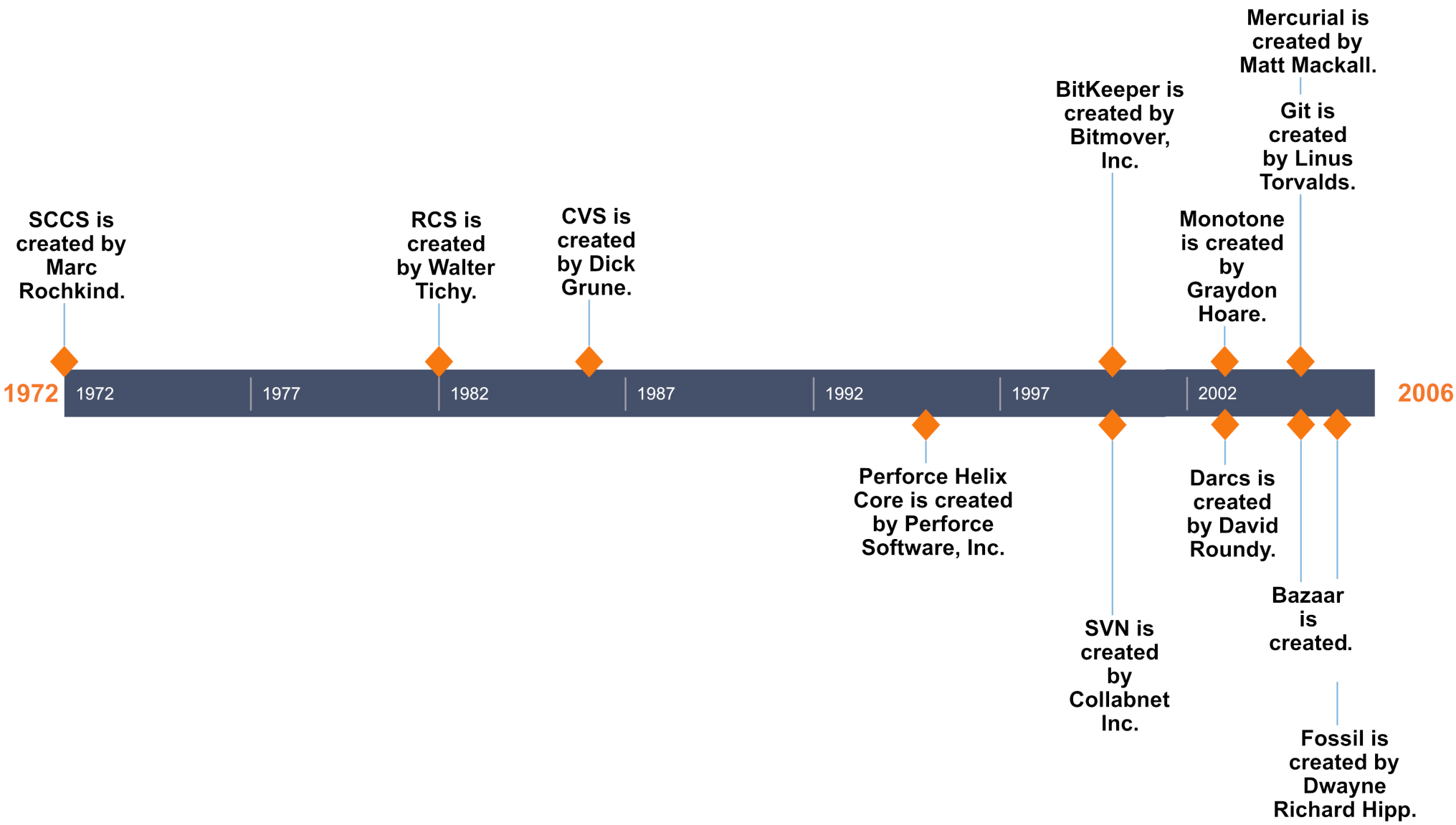
Version control systems are a category of software tools that helps in recording changes made to files by keeping a track of modifications done to the code.



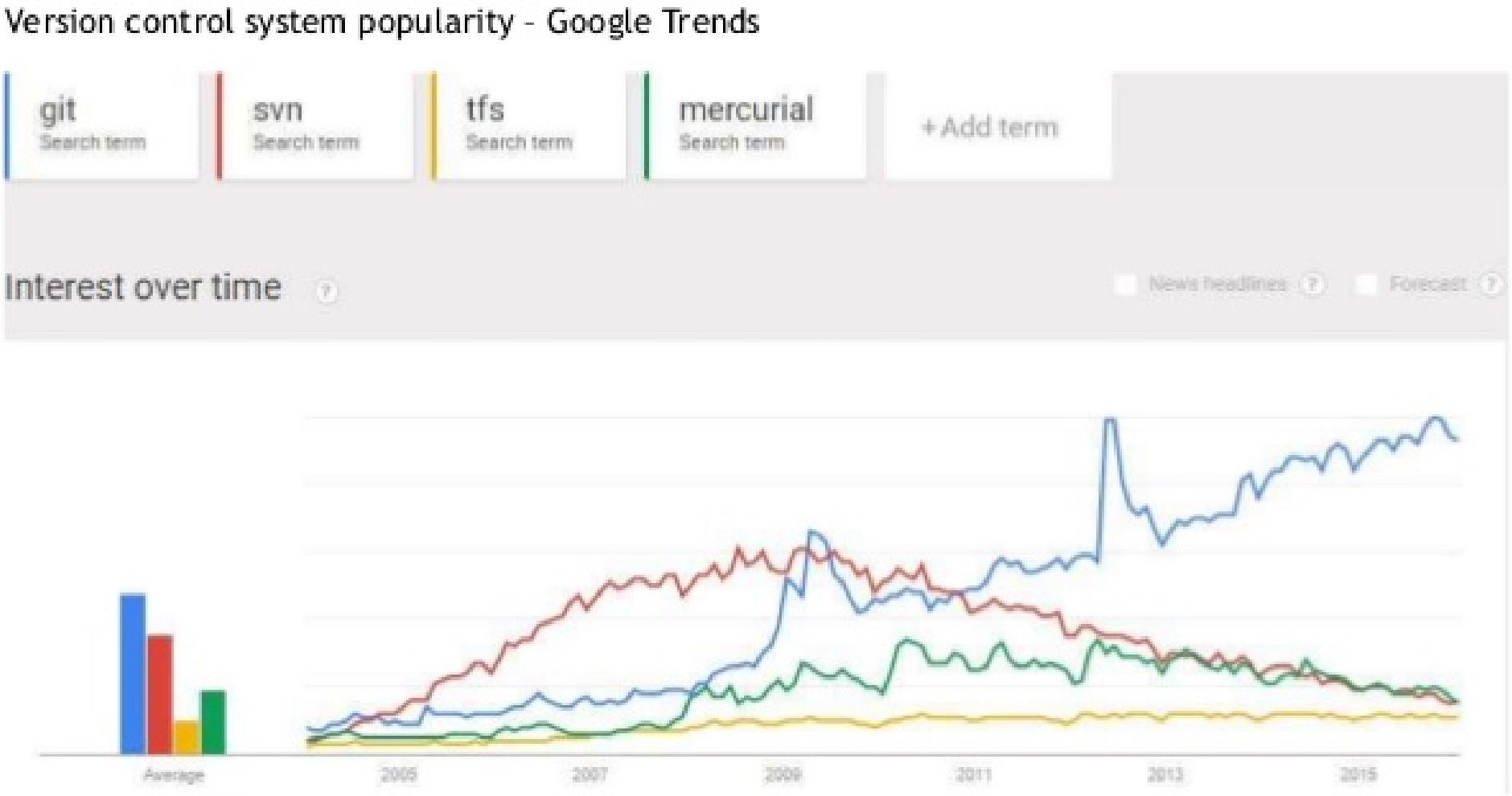

Created by Bob
Edited by Alice
Edited by by John
Approved by Manager
Centralized Repository
Contract.doc (Version 1)
Contract.doc (Version 2)
Contract.doc (Version 3)
Request for approval
Contract.doc (Final Version)
Version Notes: "I created a new contract"
Version Notes: "I edited section 3"
Version Notes: "I added a new section"
Version Notes: "This is ready for the client"

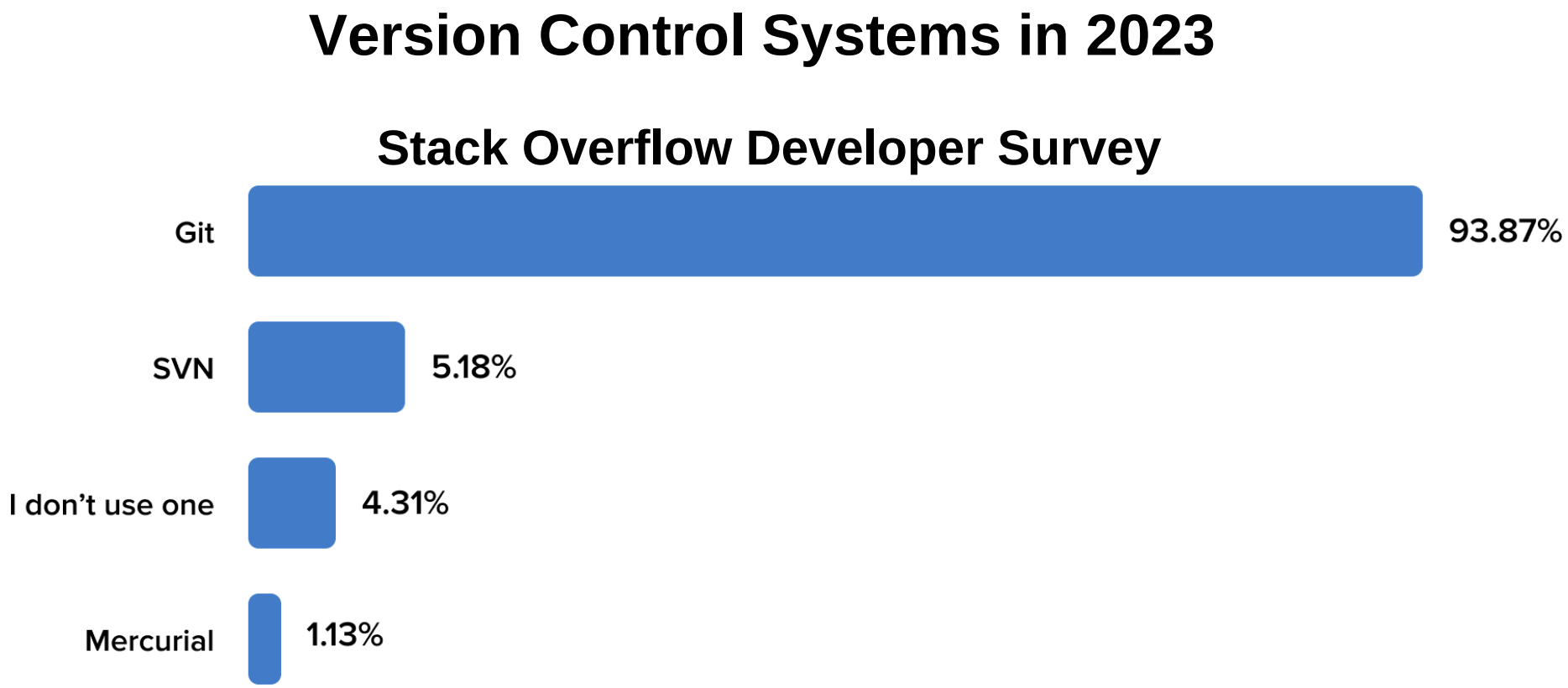
Version control systems

ADP



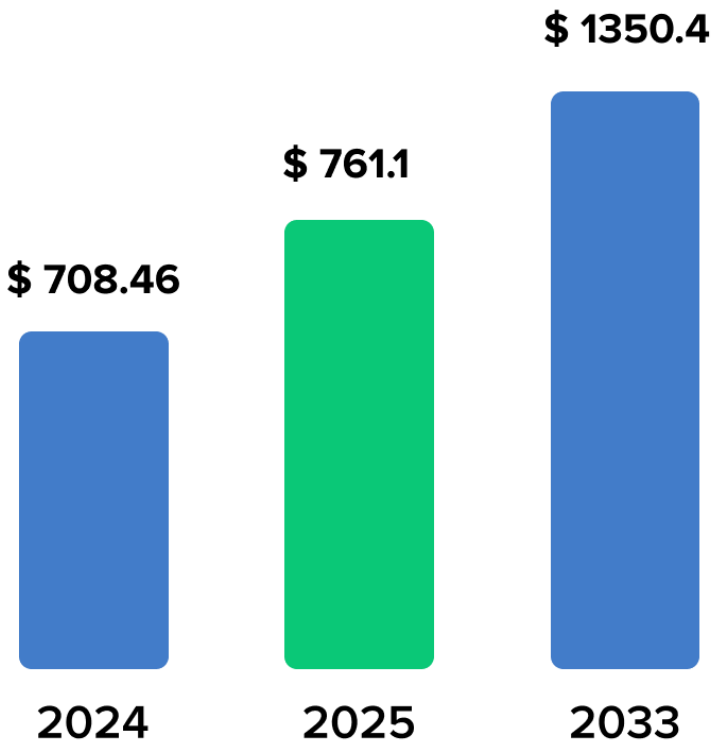
TOP VERSION CONTROL SYSTEMS


Version control system popularity - Google Trends
git
Search term
svn
Search term
tfs
Search term
mercurial
Search term
+ Add term
Interest over time
News headlines
Forecast
Average
2005
2007
2009
2011
2013
2015



https://stackoverflow.blog/2023/01/09/beyond-git-the-other-version-control-systems-developers-use/

Version Control System Market

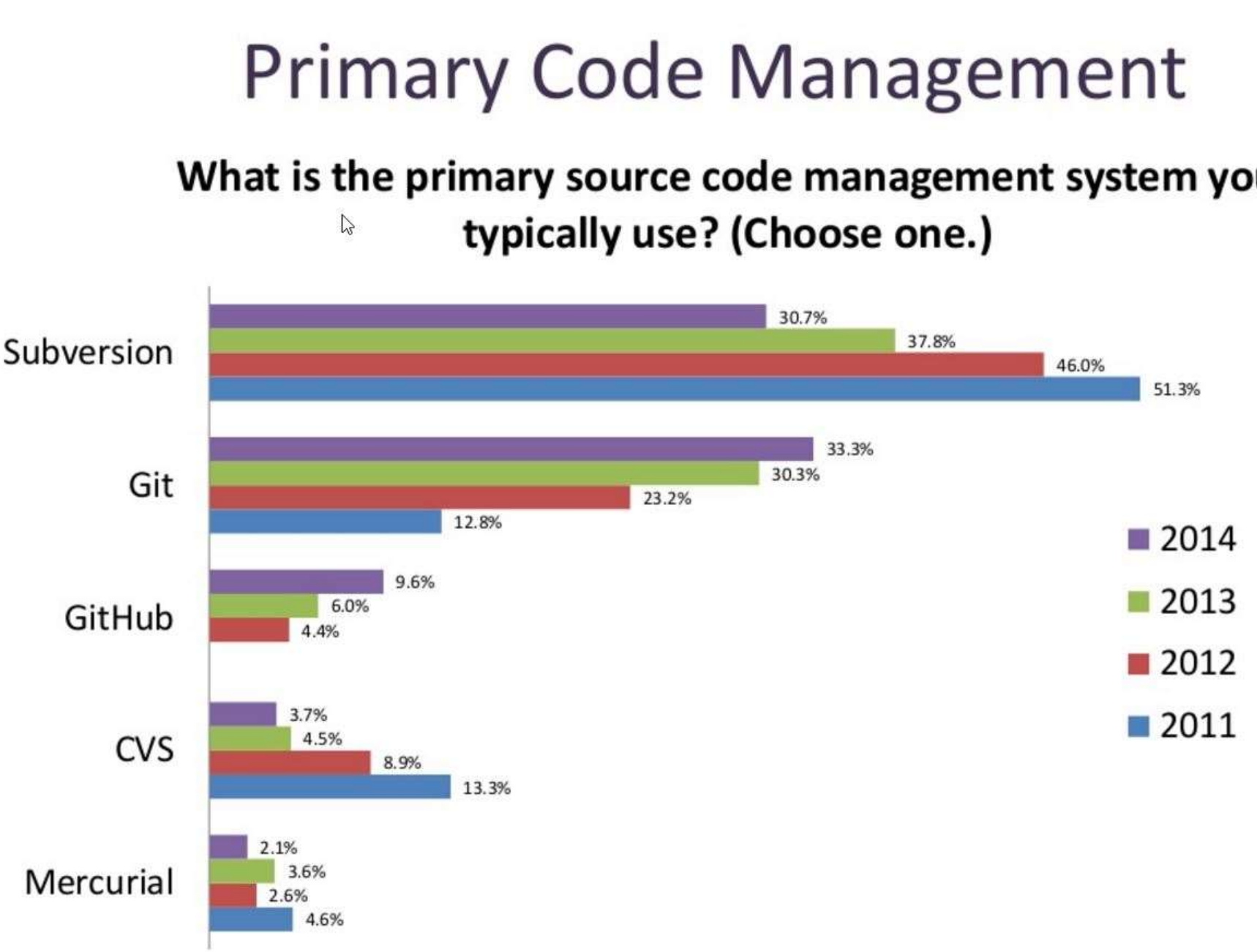


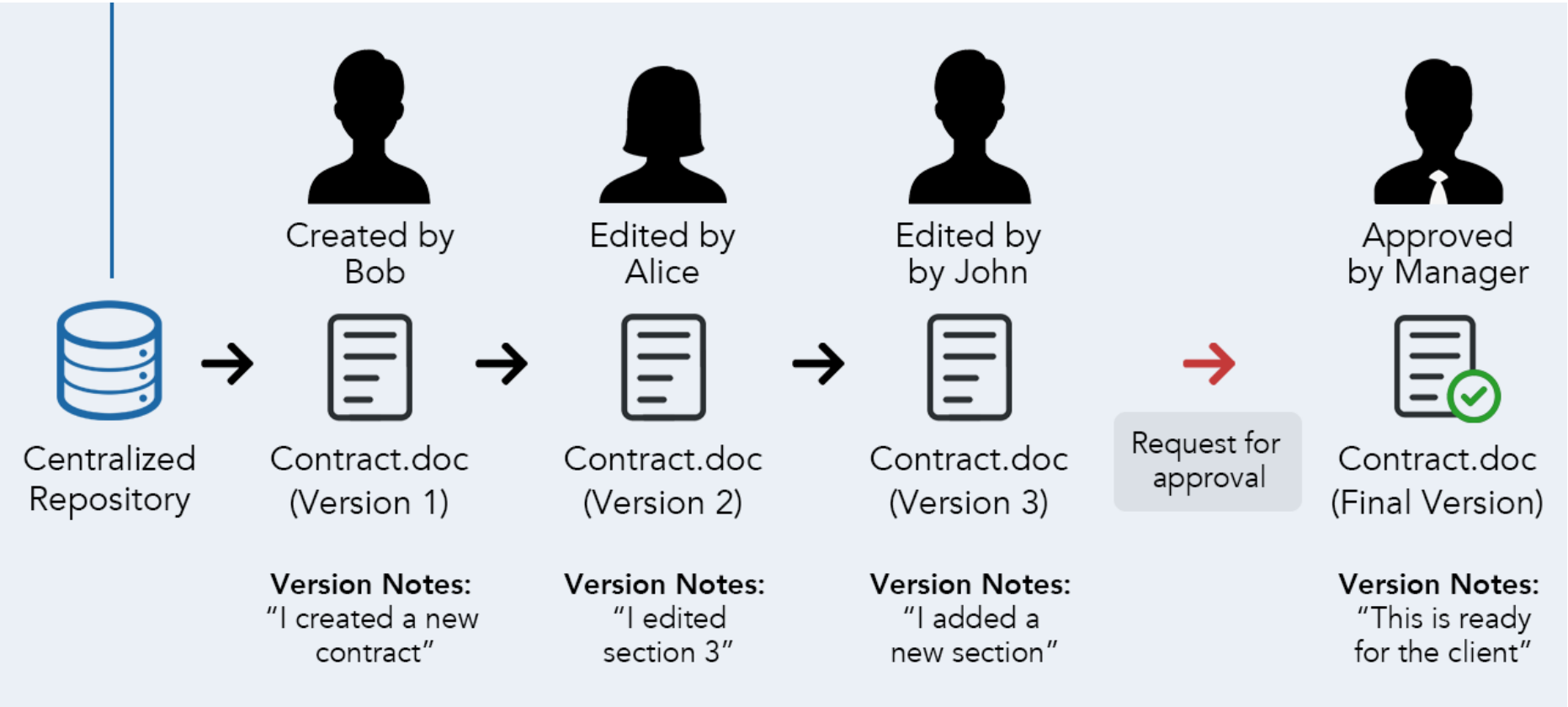
Report Insights

USD 708.46 Million	Forecast Period	USD 1350.4 Million
Base Year	2025 - 2033	Forecast Year
Market Size		Market Size

CAGR 7.43%

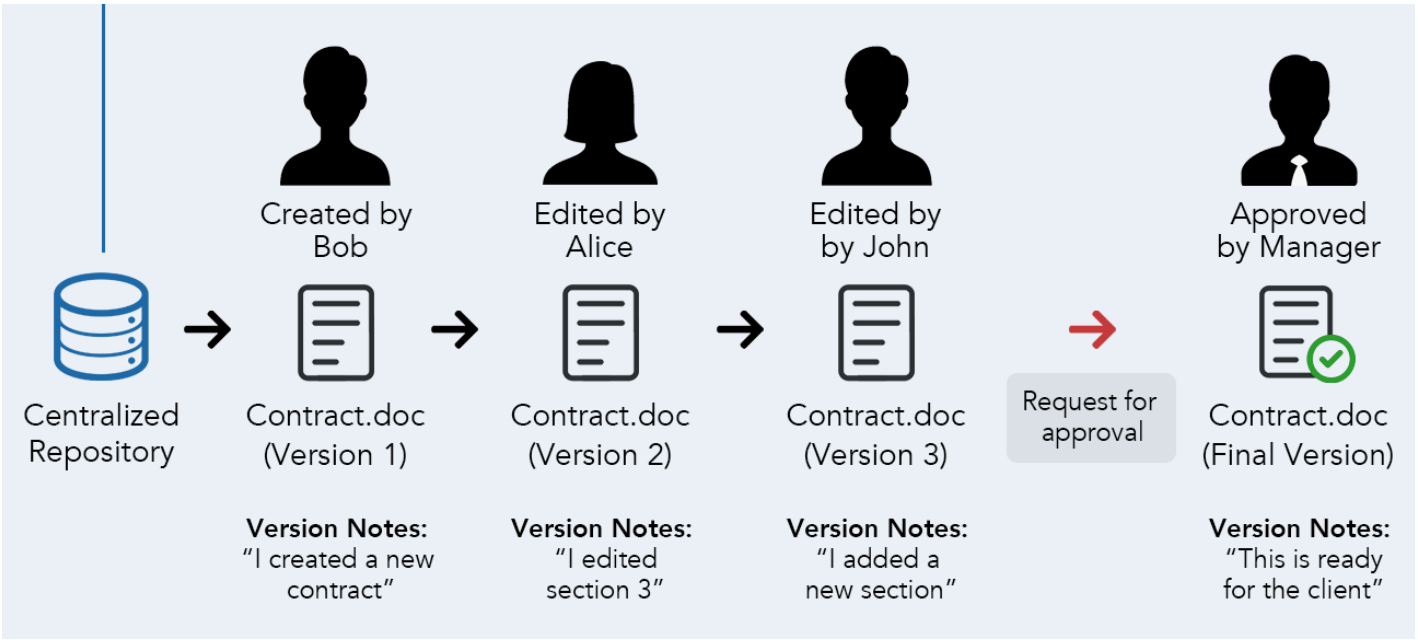
Primary Code Management

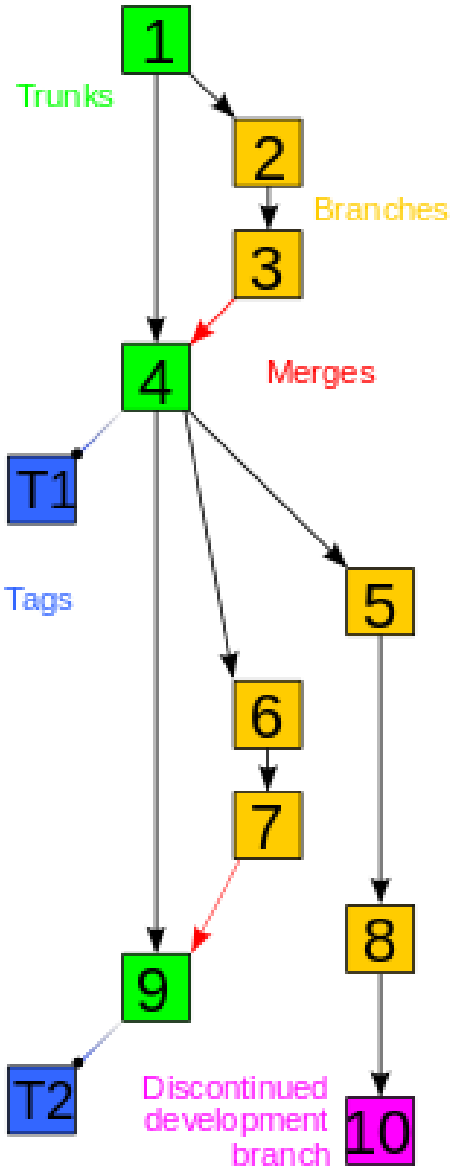


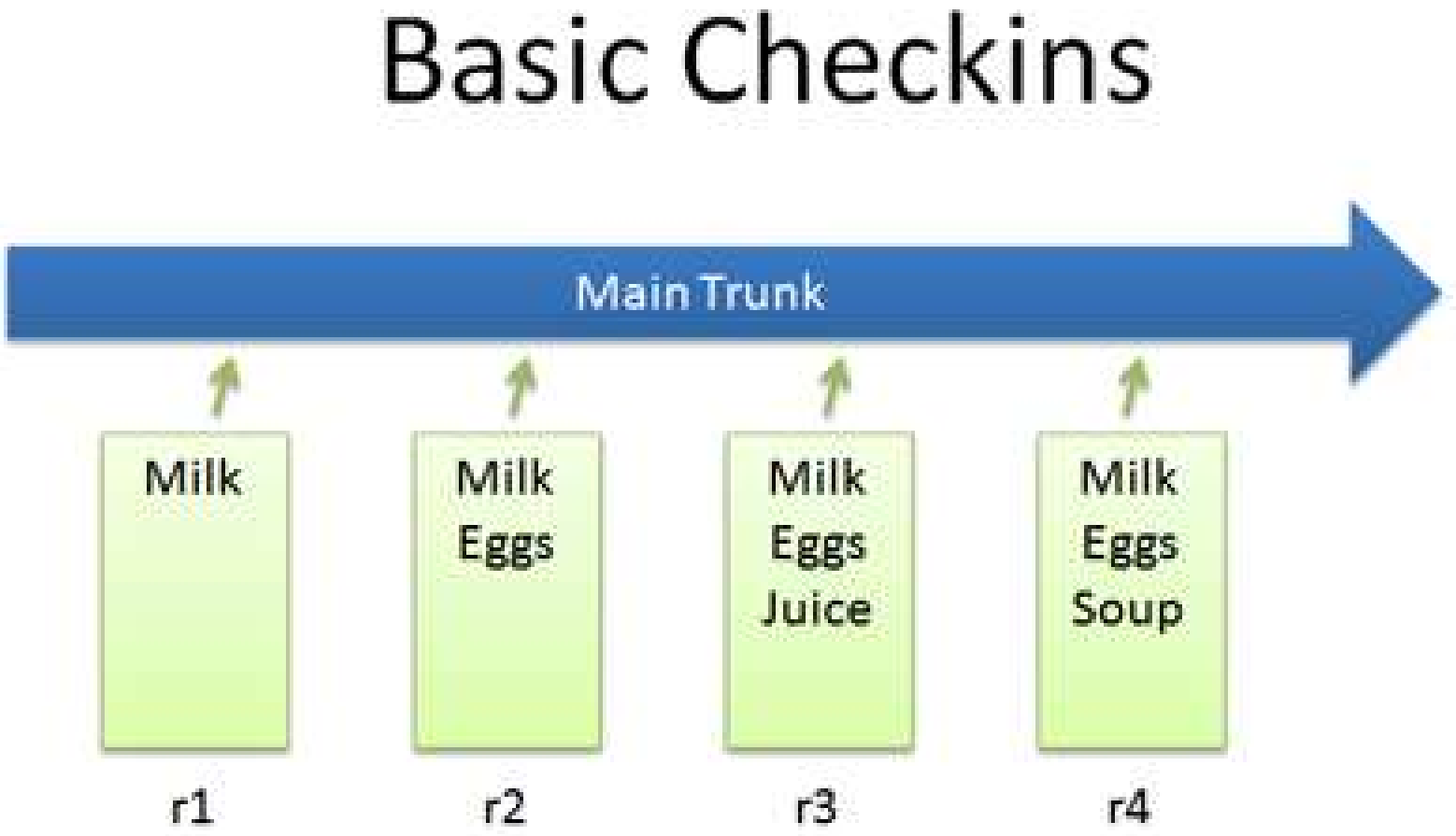

Created by Bob
Edited by Alice
Edited by by John
Approved by Manager
Centralized Repository
Contract.doc (Version 1)
Contract.doc (Version 2)
Contract.doc (Version 3)
Request for approval
Contract.doc (Final Version)
Version Notes: "I created a new contract"
Version Notes: "I edited section 3"
Version Notes: "I added a new section"
Version Notes: "This is ready for the client"

- **Branch** : A set of files under version control may be branched or forked at a point in time so that, from that time forward, two copies of those files may develop at different speeds or in different ways independently of each other.
- **Change** : A change (or diff, or delta) represents a specific modification to a document under version control.
- **Checkout** : To check out (or co) is to create a local working copy from the repository.
- **Clone** : Cloning means creating a repository containing the revisions from another repository.
- **Commit** : To commit (check in, ci or, more rarely, install, submit or record) is to write or merge the changes made in the working copy back to the repository.
- **Conflict** : A conflict occurs when different parties make changes to the same document, and the system is unable to reconcile the changes
- **Head** : Also sometimes called tip, this refers to the most recent commit, either to the trunk or to a branch
- **Merge** : A merge or integration is an operation in which two sets of changes are applied to a file or set of files
- **Pull, push** : Copy revisions from one repository into another
- **Repository** : The repository is where files' current and historical data are stored, often on a server
- **Tag** ; A tag or label refers to an important snapshot in time, consistent across many files
- **Trunk** : The unique line of development that is not a branch (sometimes also called Baseline, Mainline or Master)
- Working copy : The working copy is the local copy of files from a repository, at a specific time or revision

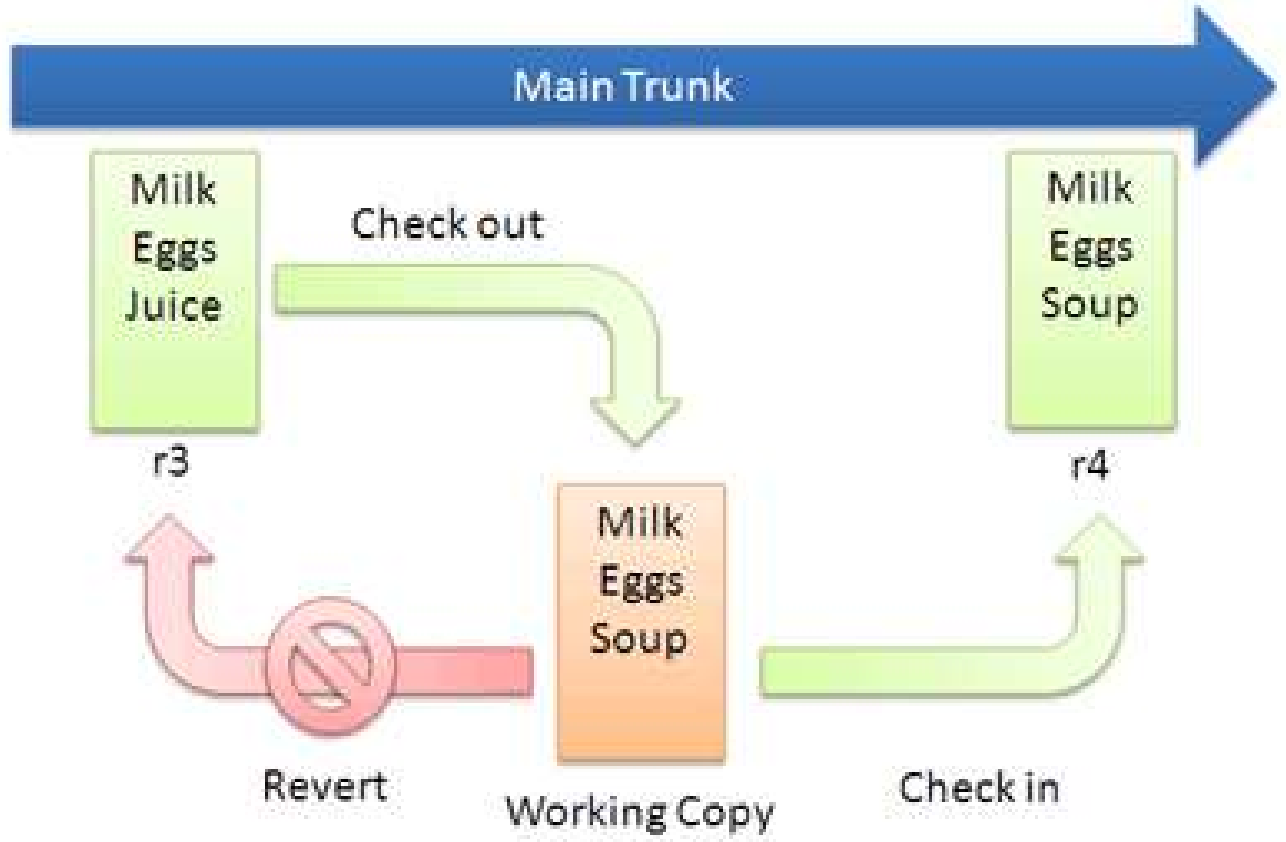
Version control systems



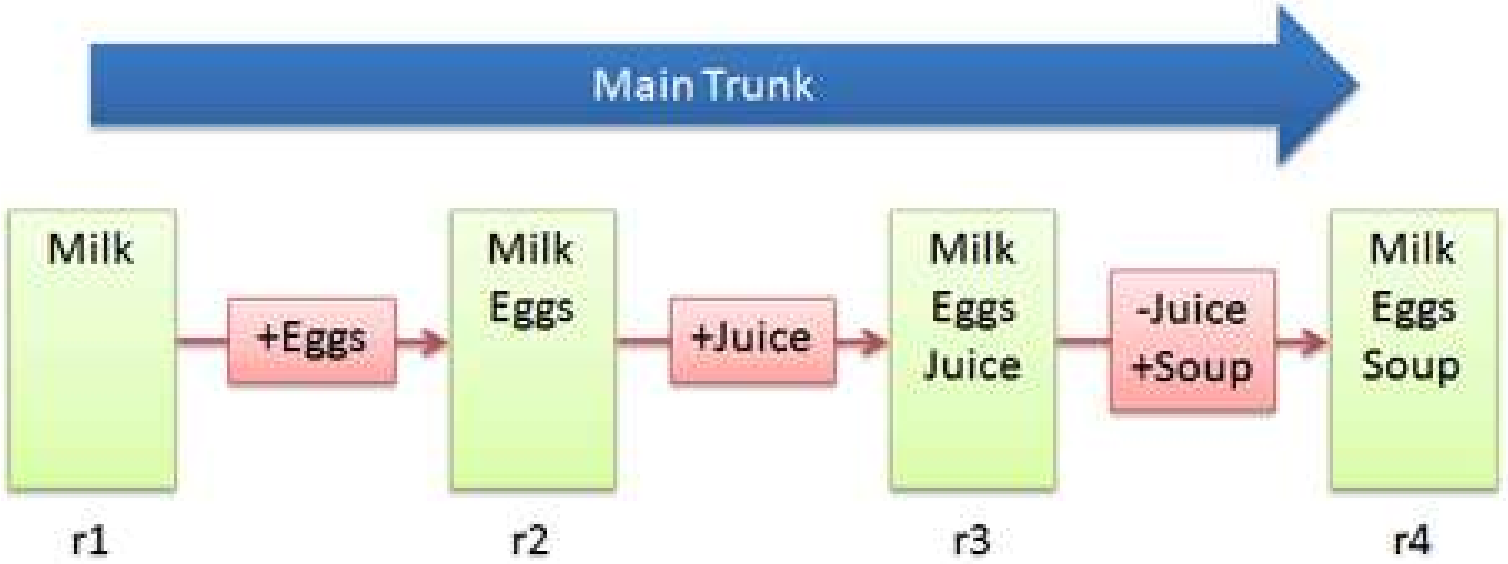


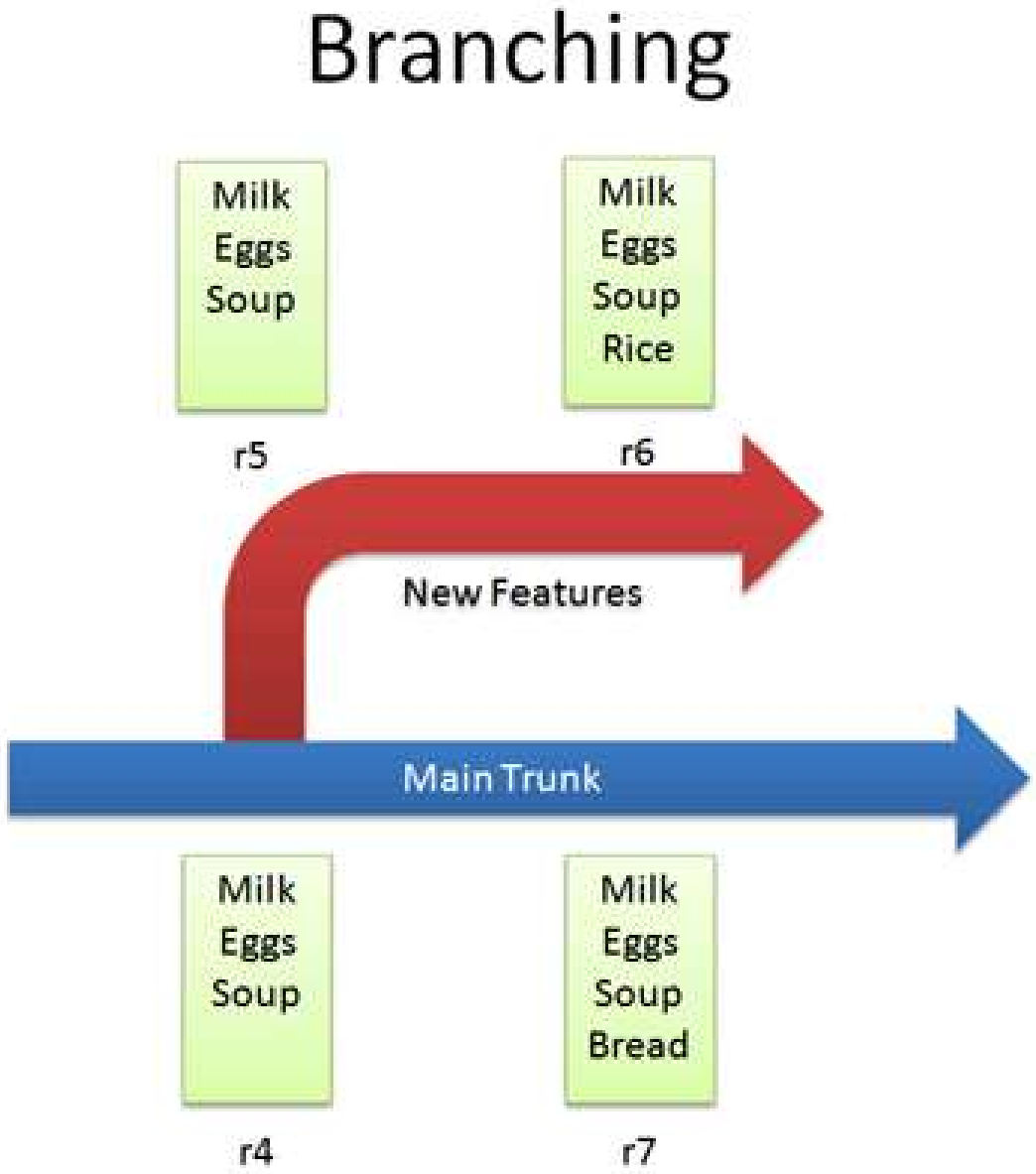

Basic Checkins
Main Trunk
Milk
Milk
Eggs
Milk
Eggs
Juice
Milk
Eggs
Soup
r1
r2
r3
r4

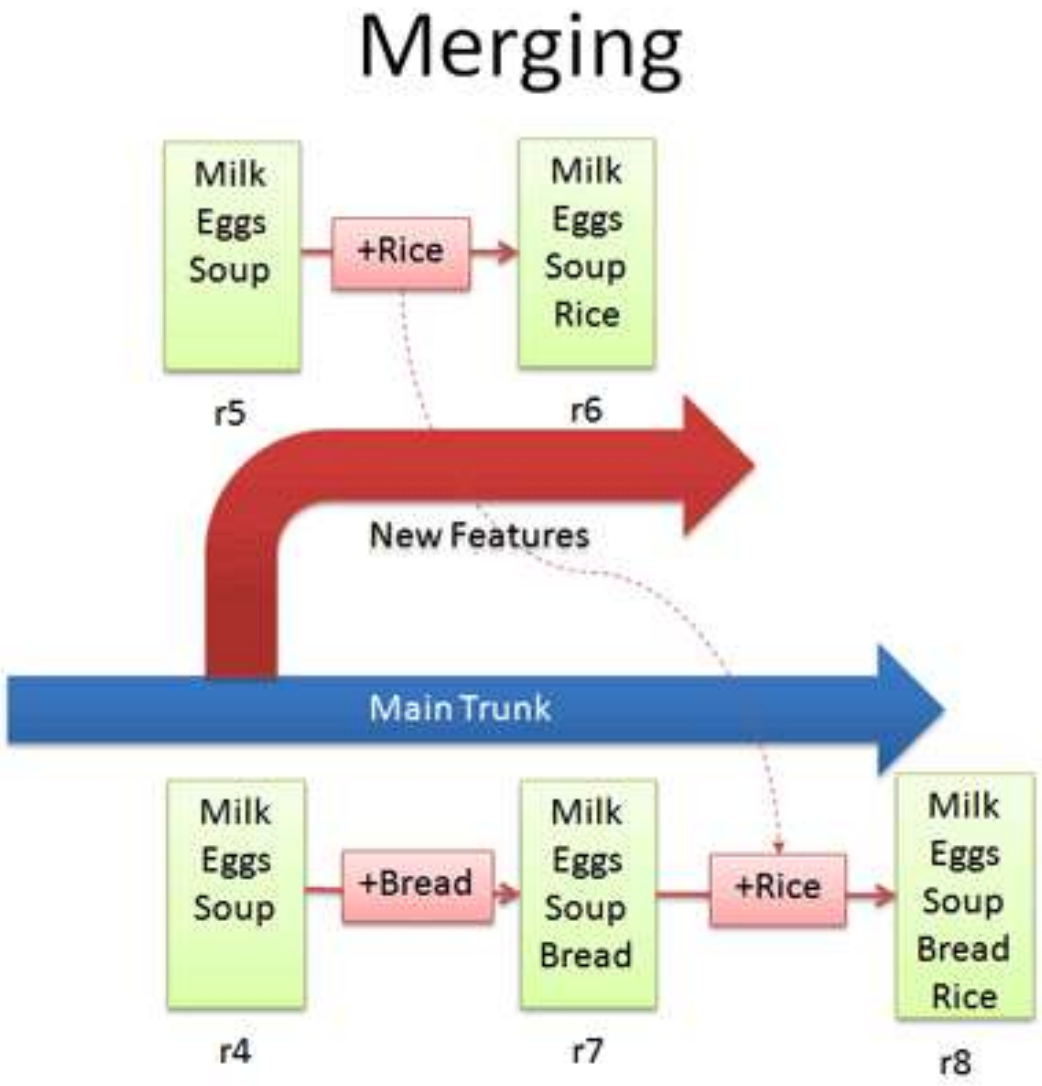
Checkout and Edit

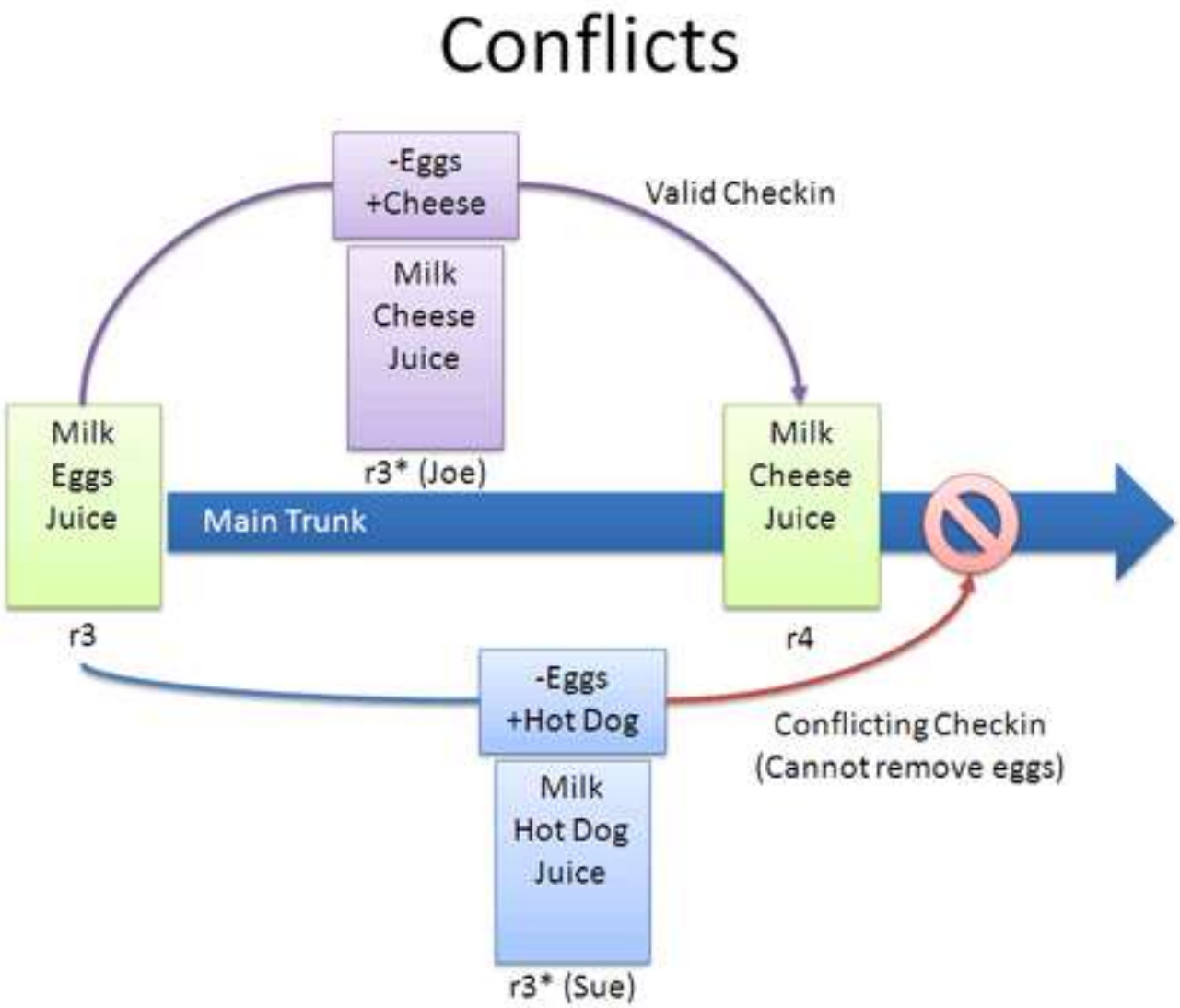


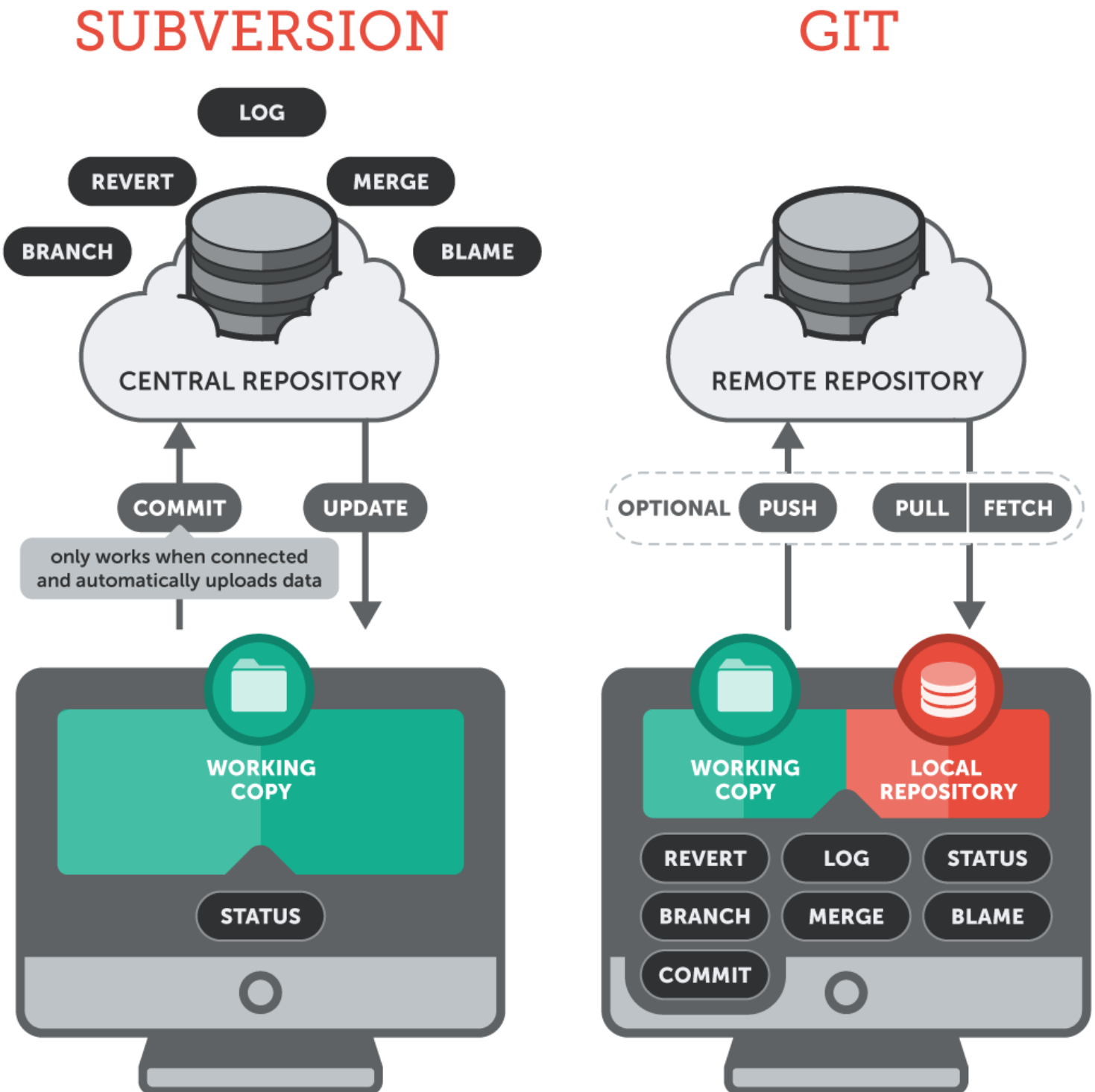
Basic Diffs


Main Trunk
Milk
+Eggs
Milk
Eggs
+Juice
Milk
Eggs
Juice
-Juice
+Soup
Milk
Eggs
Soup
r1
r2
r3
r4

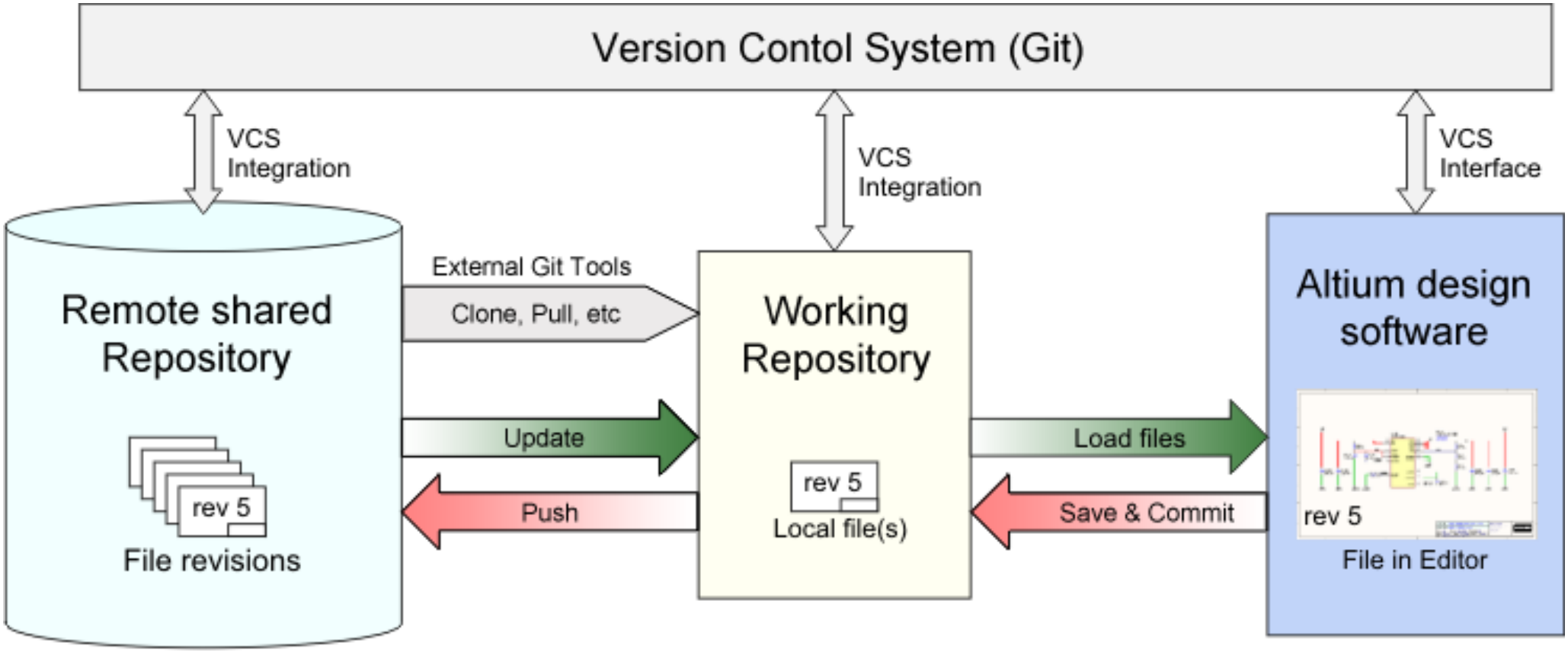

Branching
Milk
Eggs
Soup
r5
Milk
Eggs
Soup
Rice
r6
New Features
Main Trunk
Milk
Eggs
Soup
r4
Milk
Eggs
Soup
Bread
r7


Merging
Milk
Eggs
Soup
+Rice
Milk
Eggs
Soup
Rice
r5
r6
New Features
Main Trunk
Milk
Eggs
Soup
+Bread
Milk
Eggs
Soup
Bread
+Rice
Milk
Eggs
Soup
Bread
Rice
r4
r7
r8


Conflicts
-Eggs
+Cheese
Valid Checkin
Milk
Cheese
Juice
Milk
Eggs
Juice
r3* (Joe)
Main Trunk
Milk
Cheese
Juice
r3
r4
-Eggs
+Hot Dog
Conflicting Checkin
(Cannot remove eggs)
Milk
Hot Dog
Juice
r3* (Sue)


SUBVERSION
LOG
REVERT
MERGE
BRANCH
BLAME
CENTRAL REPOSITORY
COMMIT
UPDATE
only works when connected
and automatically uploads data
WORKING
COPY
STATUS
GIT
REMOTE REPOSITORY
OPTIONAL
PUSH
PULL
FETCH
WORKING
COPY
LOCAL
REPOSITORY
REVERT
LOG
STATUS
BRANCH
MERGE
BLAME
COMMIT

Version control systems



Use scenarios

”Save-as” version control
Basic cycle:
”MyCode.txt” => ”MyCodeA.txt” => ”MyCodeAA.txt” => ...you get the point

- Use OS permissioning to ”manage” concurrent access

- Local version control systems (RCS) a refinement of this scheme

Central repository model (e.g. SVN)
Basic cycle:
SVN update => SVN add / delete / copy / move => SVN status / diff => SVN update / resolve => SVN commit => (goto start)

Distributed repositories model (e.g. Git)
Basic cycle:
Git init => git add => git commit => git push => (git pull)

```
$ git init
Initialized empty Git repository in /tmp/tmp.IMBYSY7R8Y/.git/
$ cat > README << 'EOF'
> Git is a distributed revision control system.
> EOF
$ git add README
$ git commit
[master (root-commit) e4dcc69] You can edit locally and push
to any remote.
 1 file changed, 1 insertion(+)
 crate mode 100644 README
$ git remote add origin git@github.com:cdown/thats.git
$ git push -u origin master
```

2005

Linus Torvalds

Torvalds at LinuxCon Europe 2014

Born	Linus Benedict Torvalds 28 December 1969 (age 51) Helsinki, Finland

```
$ git init
Initialized empty Git repository in /tmp/tmp.IMBYSY7R8Y/
$ cat > README << 'EOF'
> Git is a distributed revision control system.
> EOF
$ git add README
$ git commit
[master (root-commit) e4dcc69] You can edit locally and push
to any remote.
 1 file changed, 1 insertion(+)
 crate mode 100644 README
$ git remote add origin git@github.com:cdown/thats.git
$ git push -u origin master
```


git + GitHub

GitHub

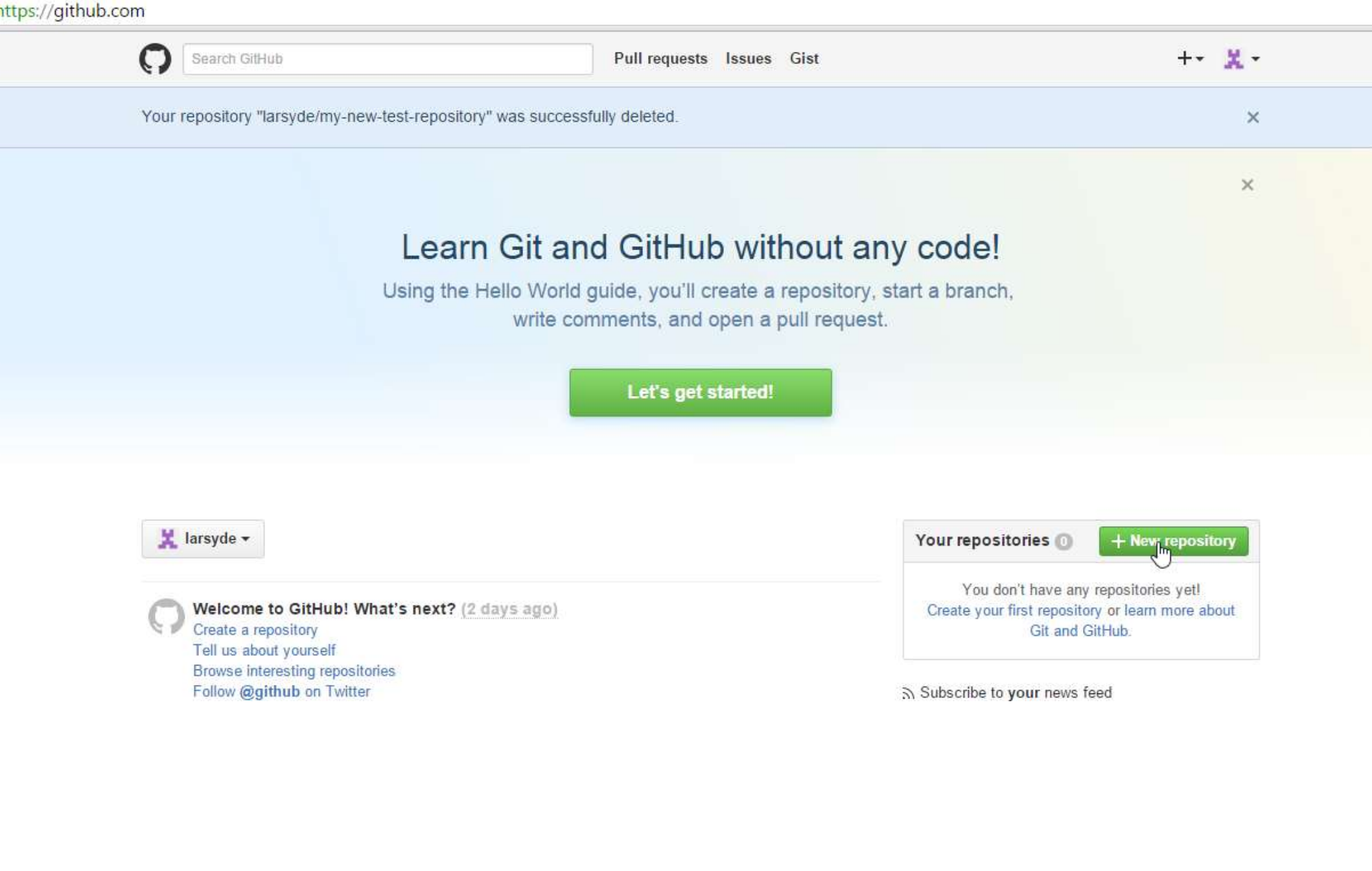
- What is it ?
  - Git
  - Git with a web interface
  - Git hosting
  - Issue and request tracking system
  - Documentation system
  - A social media platform (feeds, followers, wikis, newsletter, social graph) ?

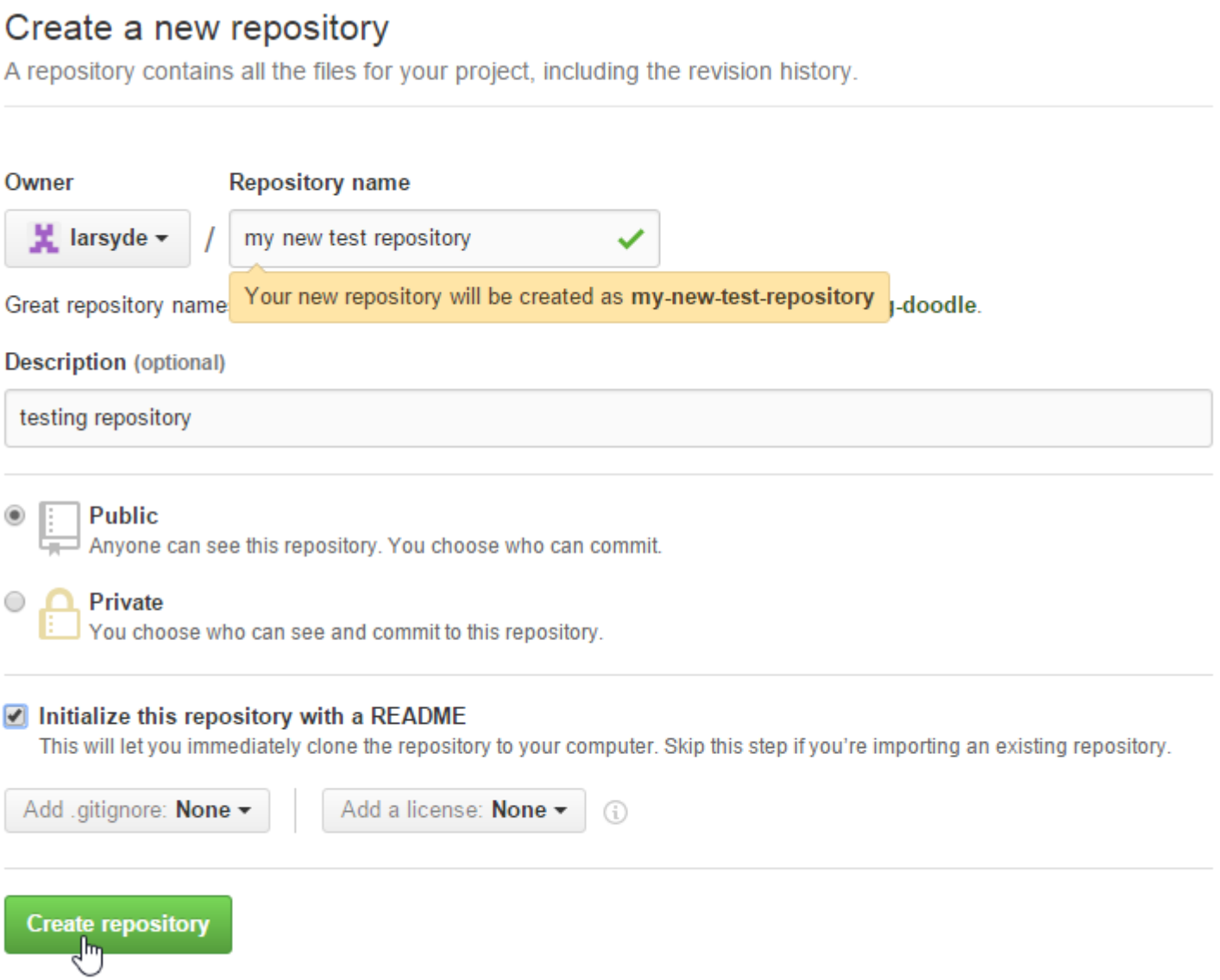
GitHub

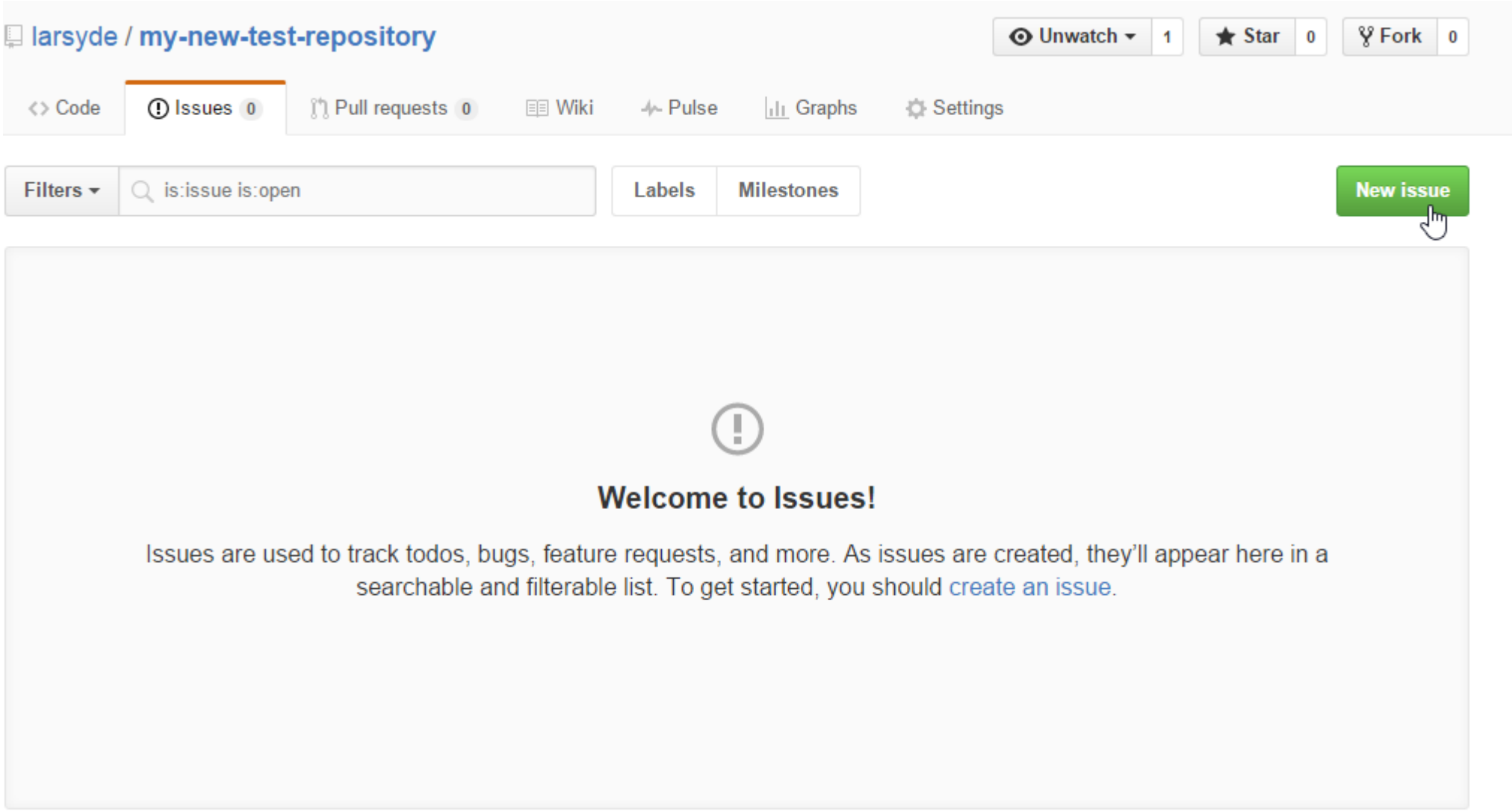
- 12M users, 31M repositories
- Why ?
  - Free(mium)
  - Natural fit to open source / non-colocated collaboration
  - Popular! (network effects, winner takes all)
  - Web interface / accessibility
- Why not ?
  - Arguably a steep learning curve
  - Maybe overkill / bad fit for certain scenarios = don't make a sheeple decision

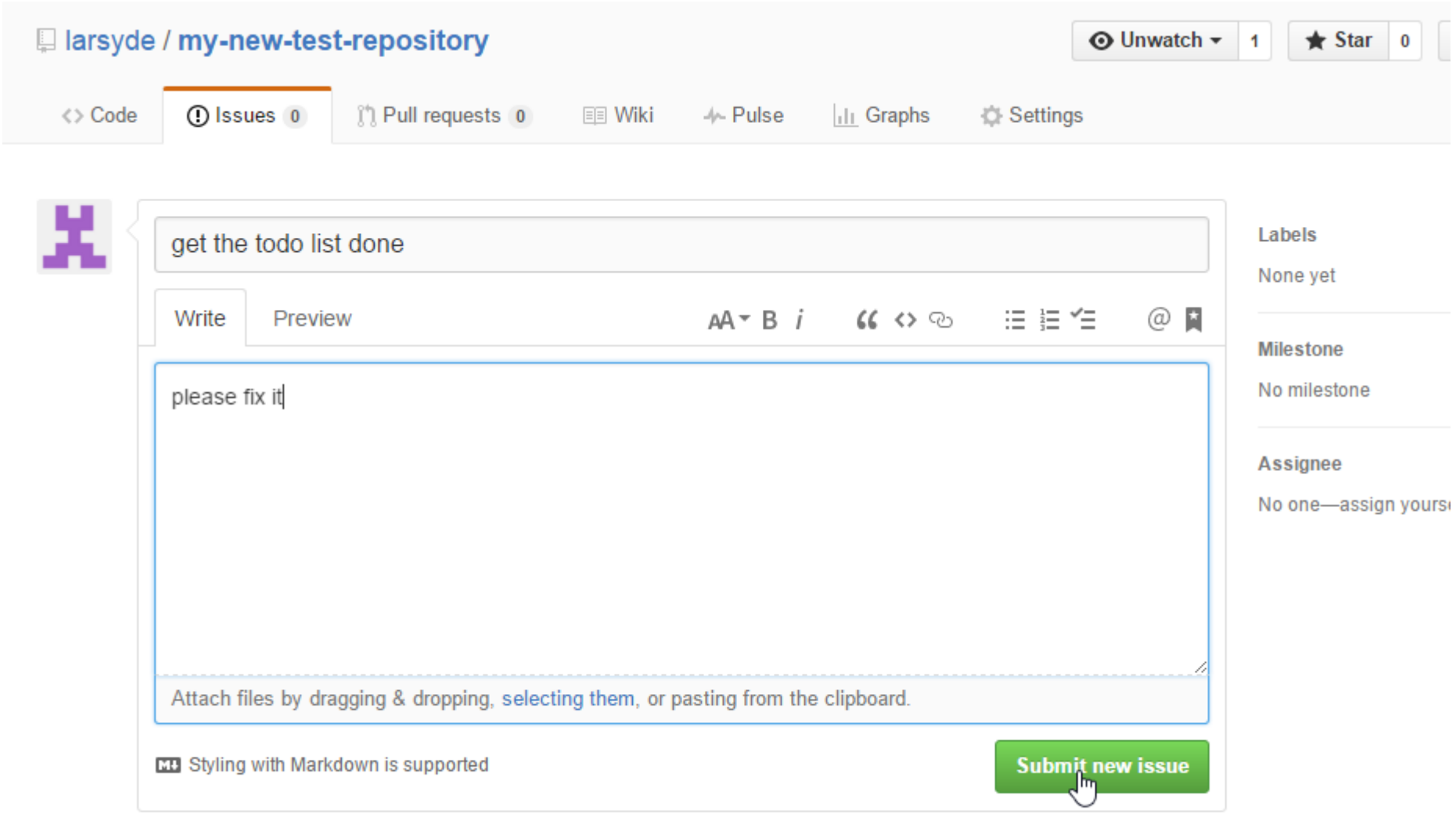
GitHub

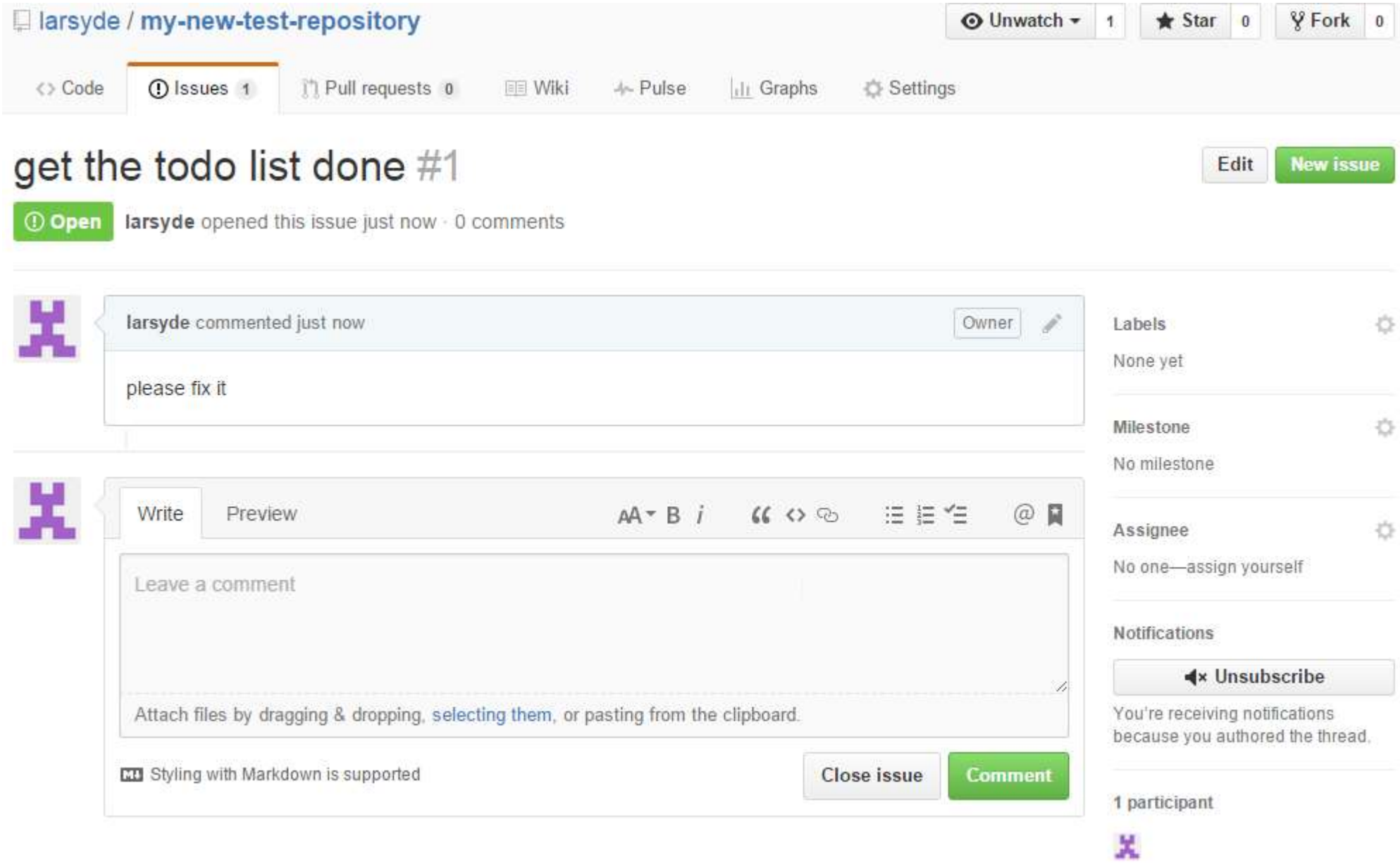
- Tooling and interfaces
  - https://git.wiki.kernel.org/index.php/InterfacesFrontendsAndTools
- GUI clients (fat clients, various platforms)
  - http://git-scm.com/downloads/guis
  - Github desktop: https://desktop.github.com/
- Branching model
  - http://nvie.com/posts/a-successful-git-branching-model/

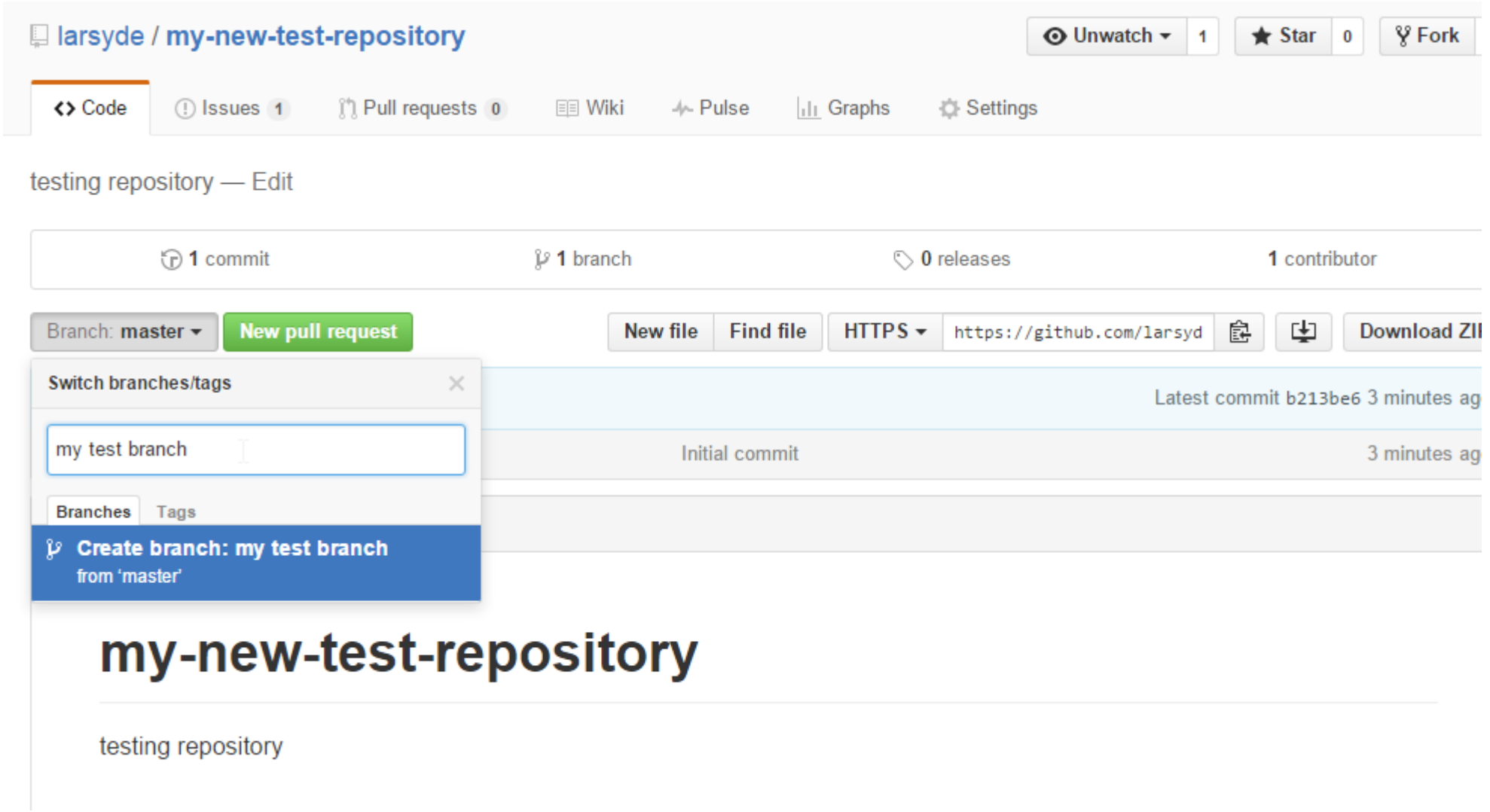

https://github.com
Search GitHub
Pull requests
Issues
Gist
Your repository "larsyde/my-new-test-repository" was successfully deleted.
Learn Git and GitHub without any code!
Using the Hello World guide, you'll create a repository, start a branch, write comments, and open a pull request.
Let's get started!
larsyde
Welcome to GitHub! What's next? (2 days ago)
Create a repository
Tell us about yourself
Browse interesting repositories
Follow @github on Twitter
Your repositories 0
+ New repository
You don't have any repositories yet!
Create your first repository or learn more about Git and GitHub.
Subscribe to your news feed

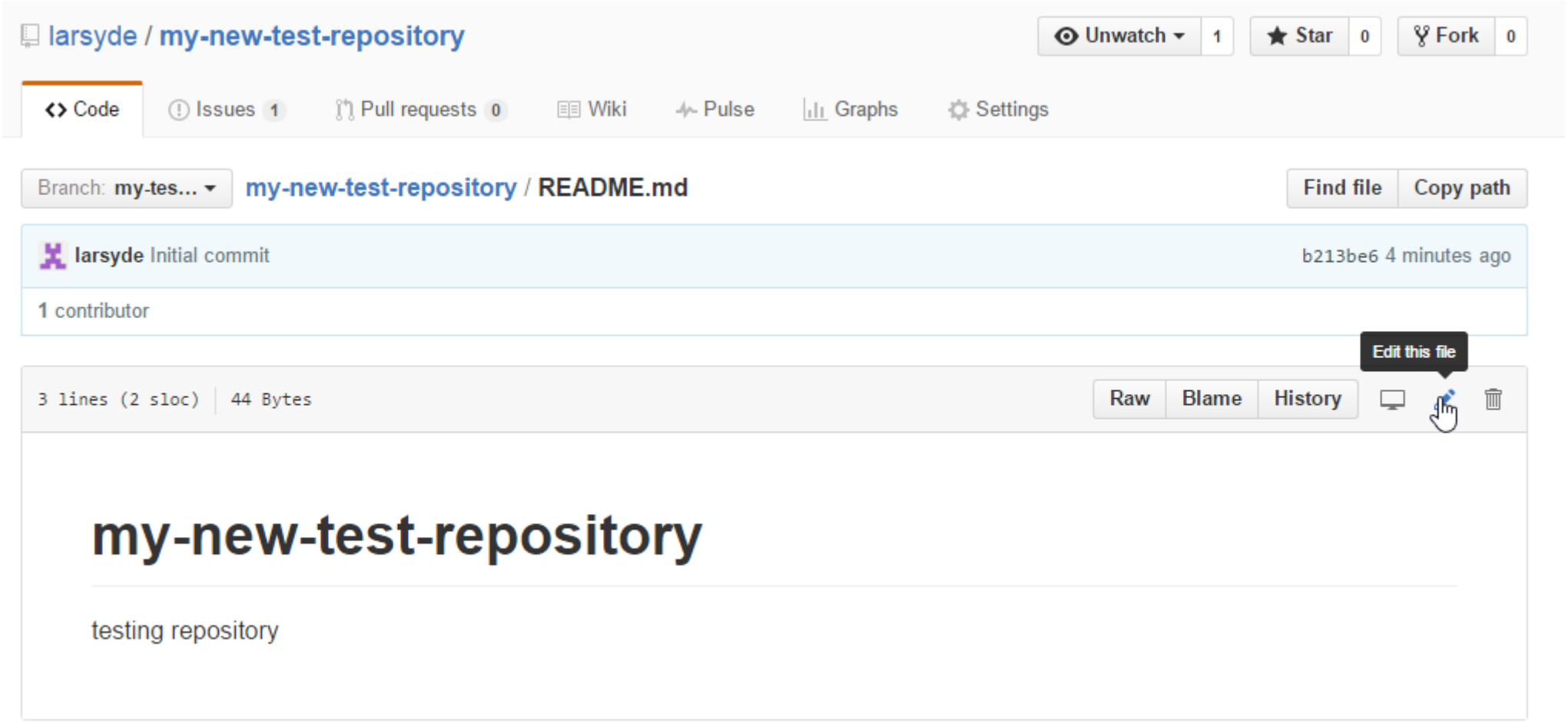

Create a new repository
A repository contains all the files for your project, including the revision history.
Owner
Repository name
larsyde
/
my new test repository
Great repository name
Your new repository will be created as my-new-test-repository
-doodle.
Description (optional)
testing repository
Public
Anyone can see this repository. You choose who can commit.
Private
You choose who can see and commit to this repository.
Initialize this repository with a README
This will let you immediately clone the repository to your computer. Skip this step if you're importing an existing repository.
Add .gitignore: None
Add a license: None
Create repository

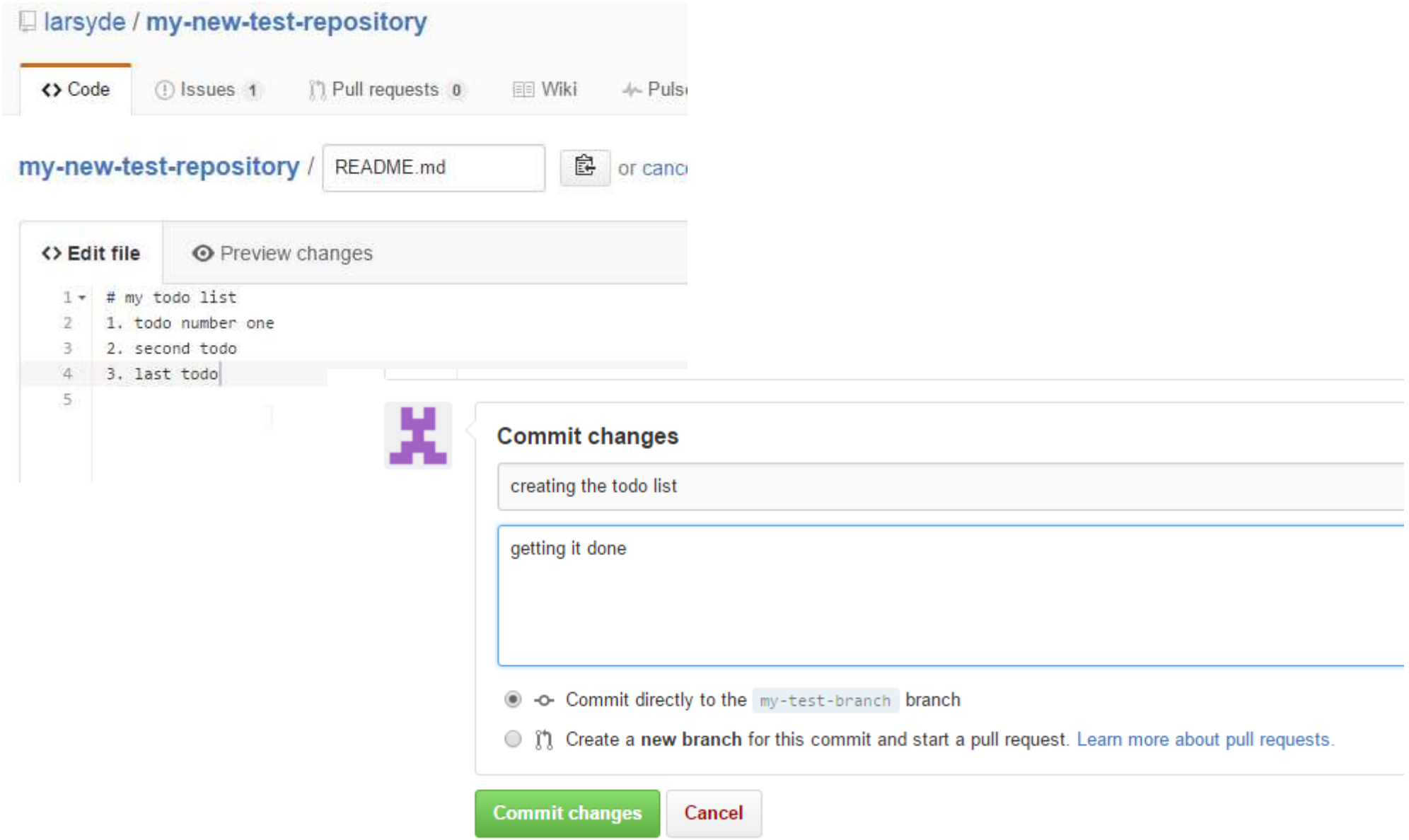

larsyde / my-new-test-repository
Unwatch 1
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Wiki
Pulse
Graphs
Settings
Filters
is:issue is:open
Labels
Milestones
New issue
Welcome to Issues!
Issues are used to track todos, bugs, feature requests, and more. As issues are created, they'll appear here in a searchable and filterable list. To get started, you should create an issue.

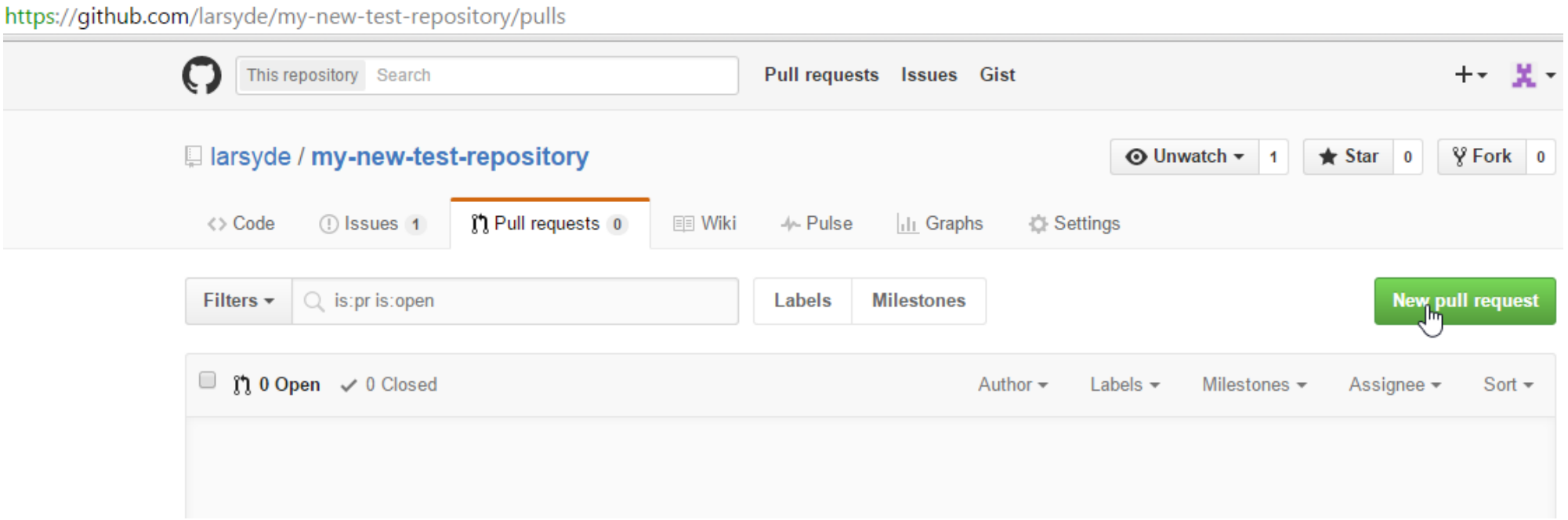

larsyde / my-new-test-repository
Unwatch 1
Star 0
Code
Issues 0
Pull requests 0
Wiki
Pulse
Graphs
Settings
get the todo list done
Write
Preview
please fix it
Attach files by dragging & dropping, selecting them, or pasting from the clipboard.
Styling with Markdown is supported
Submit new issue
Labels
None yet
Milestone
No milestone
Assignee
No one—assign yours


larsyde / my-new-test-repository
Unwatch 1
Star 0
Fork 0
Code
Issues 1
Pull requests 0
Wiki
Pulse
Graphs
Settings
get the todo list done #1
Edit
New issue
Open
larsyde opened this issue just now · 0 comments
larsyde commented just now
Owner
please fix it
Write
Preview
Leave a comment
Attach files by dragging & dropping, selecting them, or pasting from the clipboard.
Styling with Markdown is supported
Close issue
Comment
Labels
None yet
Milestone
No milestone
Assignee
No one—assign yourself
Notifications
Unsubscribe
You're receiving notifications because you authored the thread.
1 participant

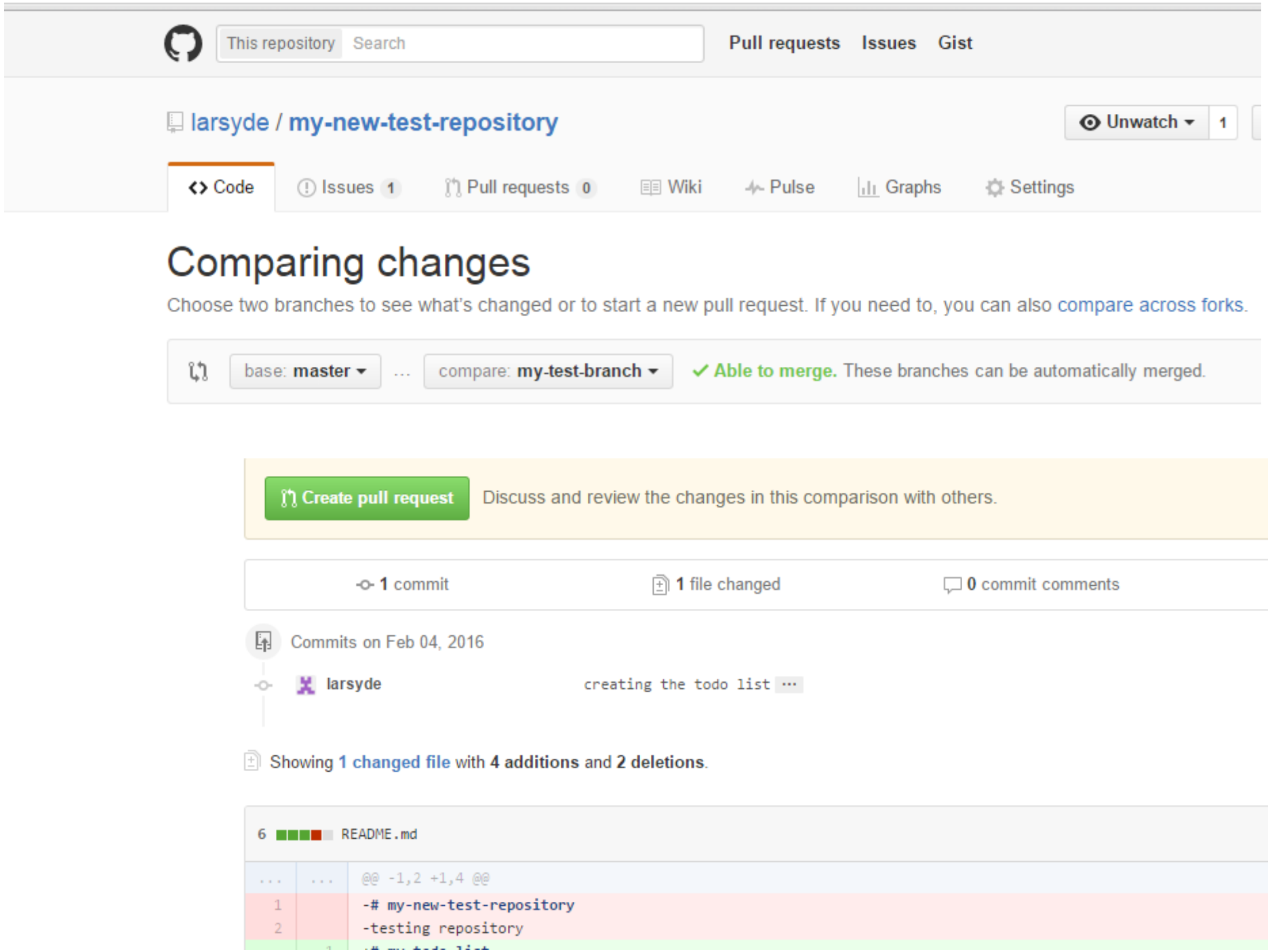

larsyde / my-new-test-repository
Unwatch 1
Star 0
Fork
Code
Issues 1
Pull requests 0
Wiki
Pulse
Graphs
Settings
testing repository — Edit
1 commit
1 branch
0 releases
1 contributor
Branch: master
New pull request
New file
Find file
HTTPS
https://github.com/larsyd
Download ZI
Switch branches/tags
my test branch
Branches
Tags
Create branch: my test branch
from 'master'
Latest commit b213be6 3 minutes ag
Initial commit
3 minutes ag
my-new-test-repository
testing repository

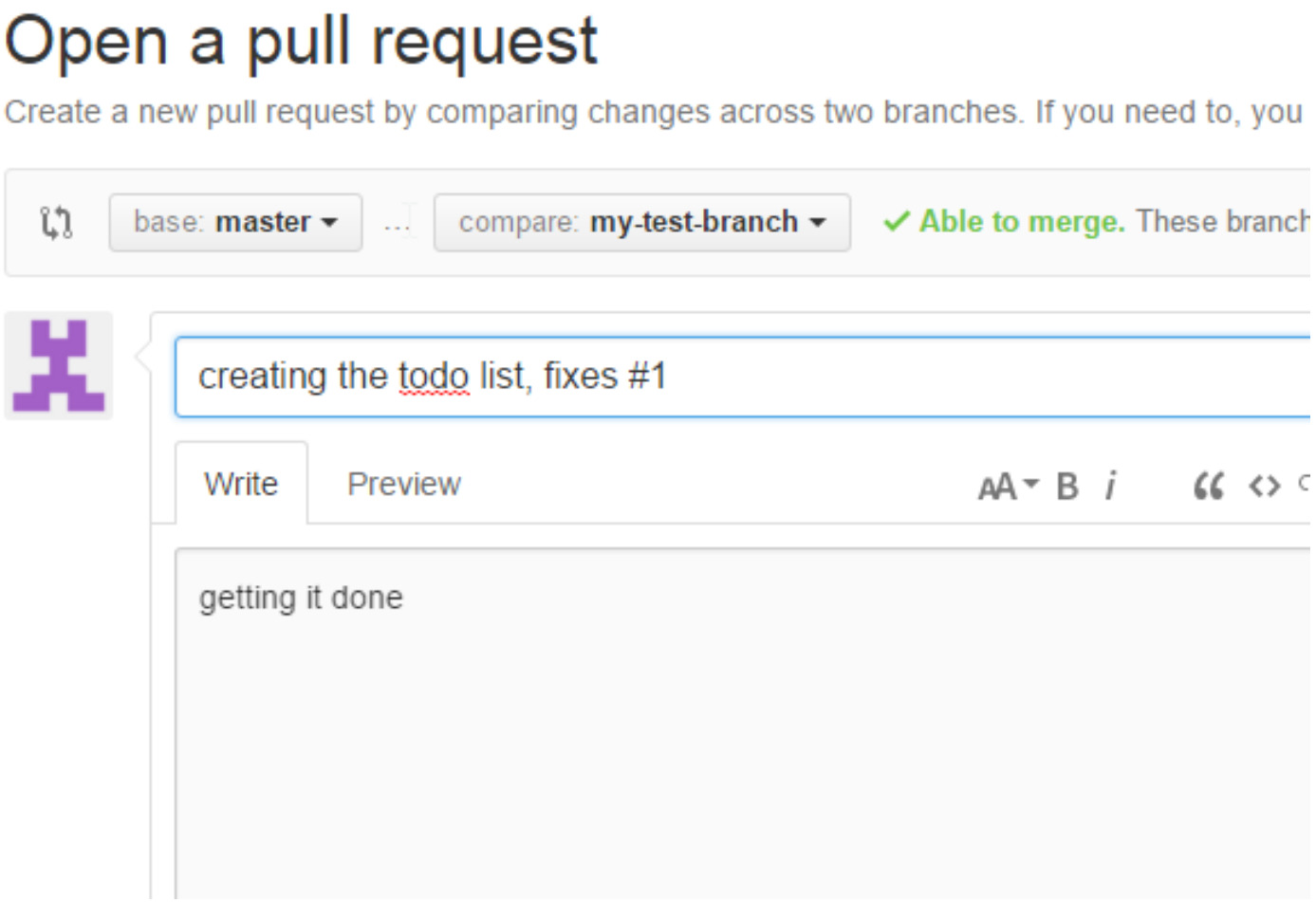

larsyde / my-new-test-repository
Unwatch 1
Star 0
Fork 0
Code
Issues 1
Pull requests 0
Wiki
Pulse
Graphs
Settings
Branch: my-tes...
my-new-test-repository / README.md
Find file
Copy path
larsyde Initial commit
b213be6 4 minutes ago
1 contributor
Edit this file
3 lines (2 sloc)
44 Bytes
Raw
Blame
History
my-new-test-repository
testing repository

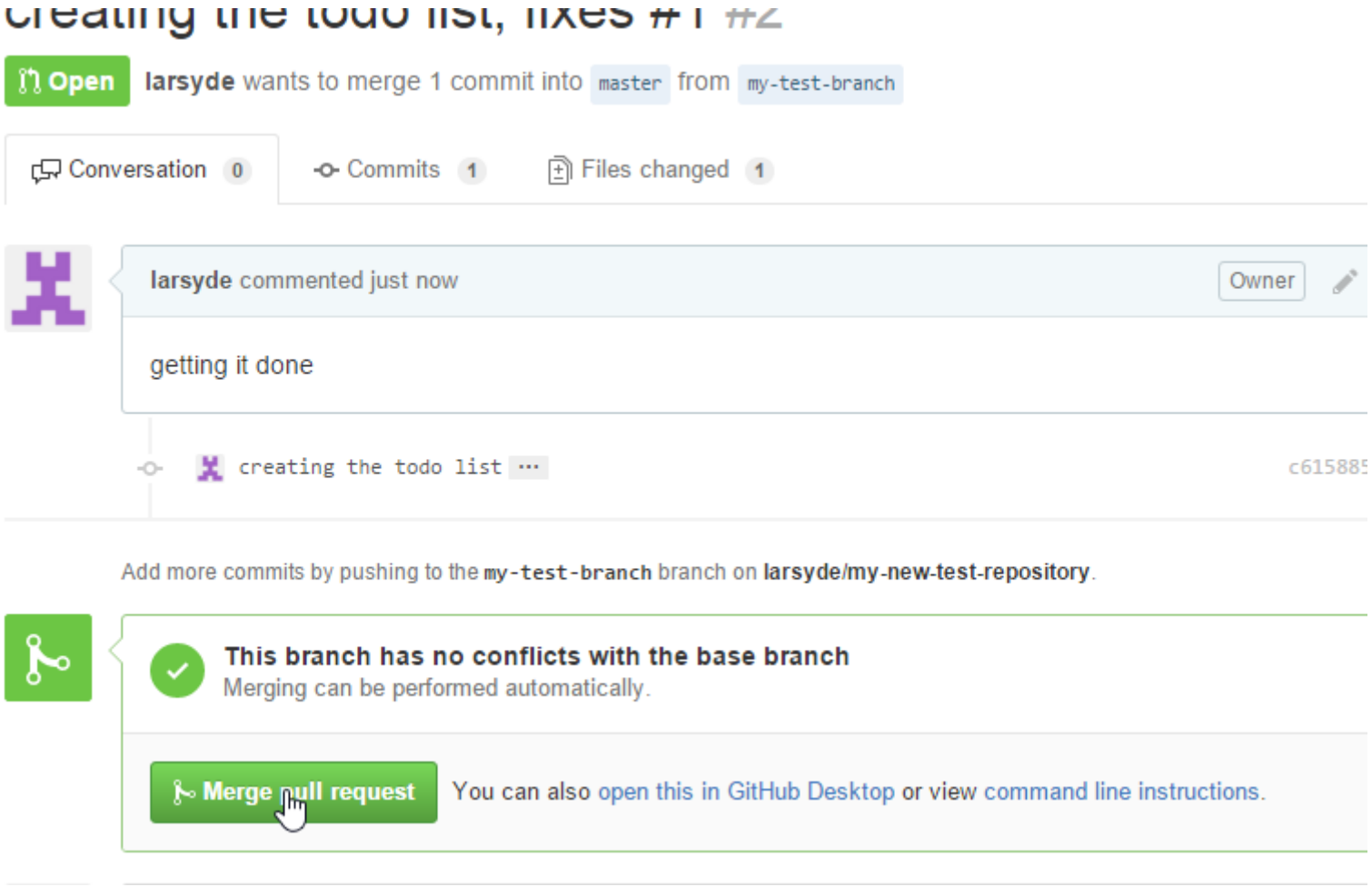

larsyde / my-new-test-repository
Code
Issues 1
Pull requests 0
Wiki
Puls
my-new-test-repository / README.md
or canc
Edit file
Preview changes
1 # my todo list
2 1. todo number one
3 2. second todo
4 3. last todo
5
Commit changes
creating the todo list
getting it done
Commit directly to the my-test-branch branch
Create a new branch for this commit and start a pull request. Learn more about pull requests.
Commit changes
Cancel


https://github.com/larsyde/my-new-test-repository/pulls
This repository
Search
Pull requests
Issues
Gist
larsyde / my-new-test-repository
Unwatch
1
Star
0
Fork
0
Code
Issues
1
Pull requests
0
Wiki
Pulse
Graphs
Settings
Filters
is:pr is:open
Labels
Milestones
New pull request
0 Open
0 Closed
Author
Labels
Milestones
Assignee
Sort

https://github.com/larsyde/my-new-test-repository/compare/master...my-test-branch




Open a pull request
Create a new pull request by comparing changes across two branches. If you need to, you
base: master
...
compare: my-test-branch
Able to merge. These branch
creating the todo list, fixes #1
Write
Preview
getting it done


creating the todo list, fixes #1 #2
Open
larsyde wants to merge 1 commit into master from my-test-branch
Conversation 0
Commits 1
Files changed 1
larsyde commented just now
Owner
getting it done
creating the todo list ···
Add more commits by pushing to the my-test-branch branch on larsyde/my-new-test-repository.
This branch has no conflicts with the base branch
Merging can be performed automatically.
Merge pull request
You can also open this in GitHub Desktop or view command line instructions.

creating the todo list, fixes #1 #2

Open **larsyde** wants to merge 1 commit into `master` from `my-test-branch`

Conversation 0 Commits 1 Files changed 1

larsyde commented a minute ago Own

getting it done

creating the todo list ···

Add more commits by pushing to the **my-test-branch** branch on **larsyde/my-new-test-repository**.

Merge pull request #2 from larsyde/my-test-branch

creating the todo list, fixes #1

Confirm merge Cancel

creating the todo list, fixes #1 #2

Merged **larsyde** merged 1 commit into `master` from `my-test-branch` just now

Conversation 0 Commits 1 Files changed 1

larsyde commented a minute ago — Owner

getting it done

`creating the todo list ···` c615885

larsyde merged commit `3bfb480` into `master` just now — Revert

Avoid bugs by automatically running your tests.

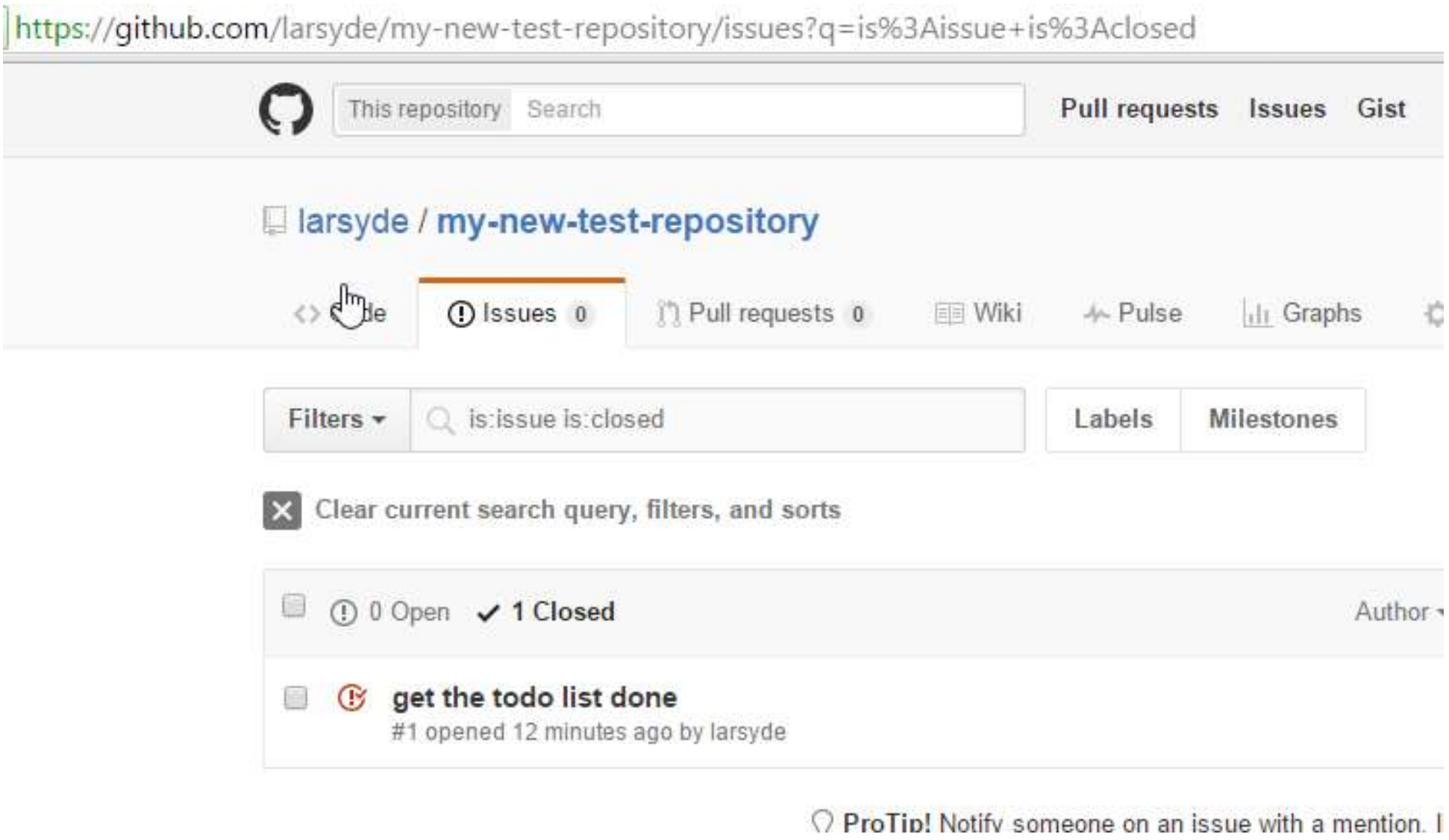
Continuous integration can help catch bugs by running your tests automatically.
Merge your code with confidence using one of our continuous integration providers.

Learn more

Pull request successfully merged and closed

You're all set—the `my-test-branch` branch can be safely deleted.

Delete branch


https://github.com/larsyde/my-new-test-repository/issues?q=is%3Aissue+is%3Aclosed
This repository
Search
Pull requests
Issues
Gist
larsyde / my-new-test-repository
Issues 0
Pull requests 0
Wiki
Pulse
Graphs
Filters
is:issue is:closed
Labels
Milestones
Clear current search query, filters, and sorts
0 Open
1 Closed
Author
get the todo list done
#1 opened 12 minutes ago by larsyde
ProTip! Notify someone on an issue with a mention. I

1. First, update your packages.

```
sudo apt update
```

2. Next, install Git and GitHub with apt-get

```
sudo apt-get install git
```

3. Finally, verify that Git is installed correctly

```
git --version
```

4. Run the following commands with your information to set a default username and email when you're going to save your work.

```
git config --global user.name "MV Thanoshan"
git config --global user.email "example@mail.com"
```

https://www.freecodecamp.org/news/the-beginners-guide-to-git-github/

Once you have the address of the repository, you need to use your terminal. Use the following command on your terminal. When you're ready you can enter this:

```
git clone [HTTPS ADDRESS]
```

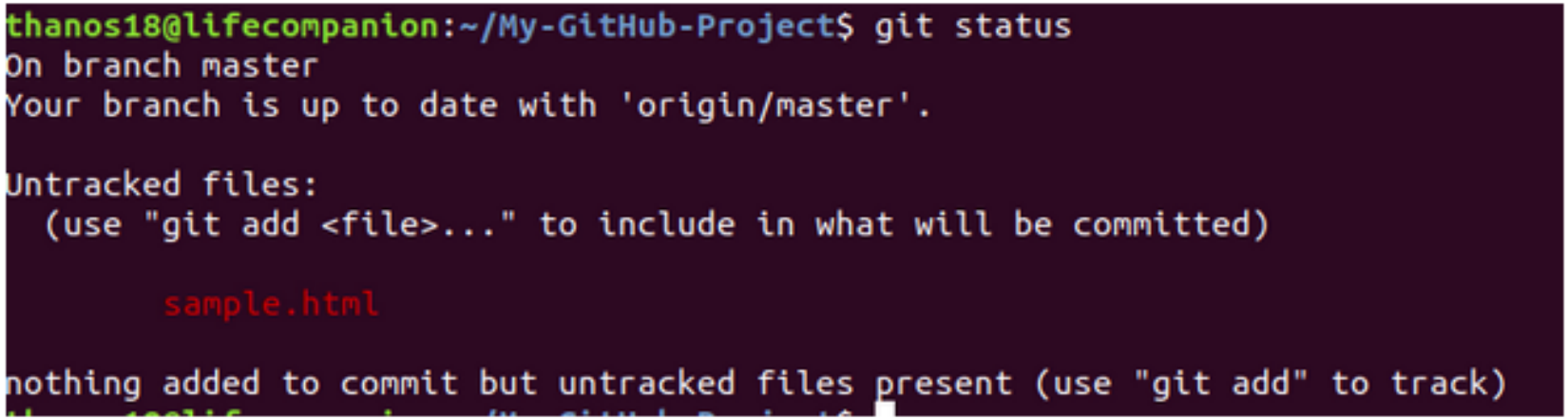
This command will make a local copy of the repository hosted at the given address.

```
Cloning into 'My-GitHub-Project'...
remote: Enumerating objects: 3, done.
remote: Counting objects: 100% (3/3), done.
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (3/3), done.
```

Output message of "git clone" command

1. “status”: The first thing you need to do is to check the files you have modified. To do this, you can type the following command to make a list of changes appear.

```
git status
```

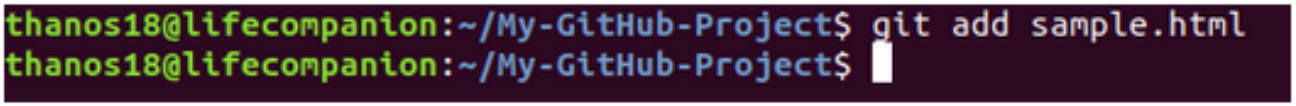


2. “add”: With the help of the change list, you can add all files you want to upload with the following command,

```
git add [FILENAME] [FILENAME] [...]
```

In our case, we’ll add a simple HTML file.

```
git add sample.html
```



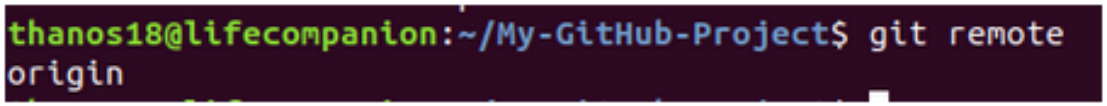
3. “commit”: Now that we have added the files of our choice, we need to write a message to explain what we have done. This message may be useful later if we want to check the change history. Here is an example of what we can put in our case.

```
git commit -m "Added sample HTML file that contain basic syntax"
```

```
thanos18@lifecompanion:~/My-GitHub-Project$  git commit -m "Added sample HTML fi
le that contain basic syntax"
[master b234227] Added sample HTML file that contain basic syntax
 1 file changed, 12 insertions(+)
 create mode 100644 sample.html
```

4. “push”: Now we can put our work on GitHub. To do that we have to ‘push’ our files to Remote. Remote is a duplicate instance of our repository that lives somewhere else on a remote server. To do this, we must know the remote’s name (Mostly remote is named origin). To figure out that name, type the following command.

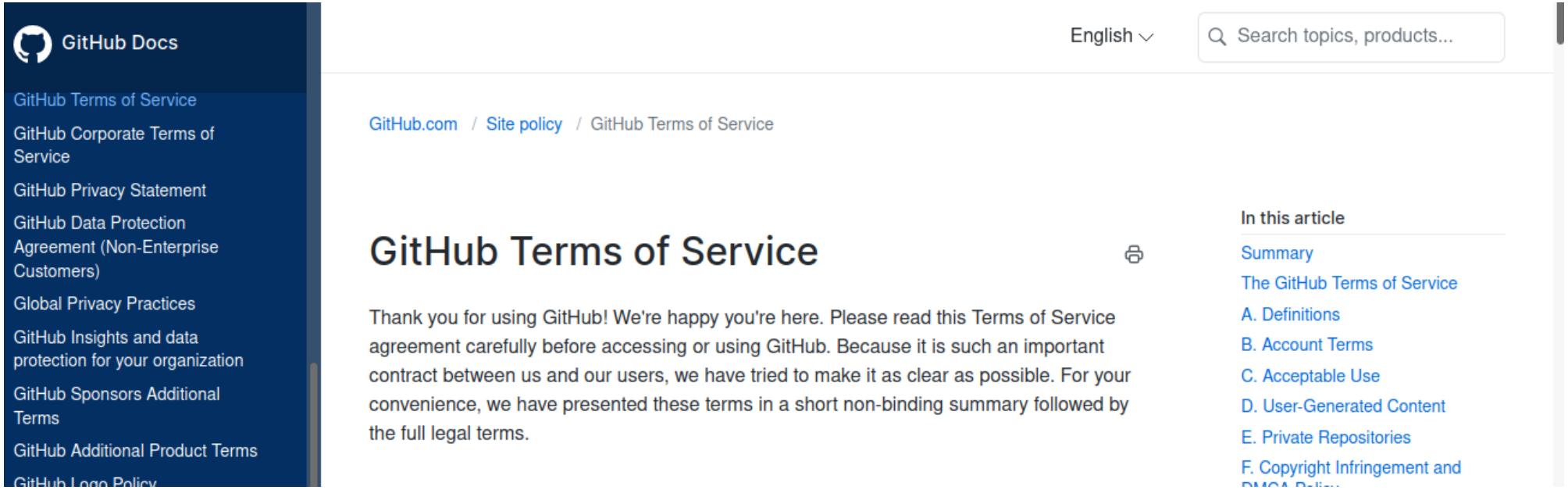
```
git remote
```



As you can see in the above image, it says that our remote’s name is origin. Now we can safely ‘push’ our work by the following command.

```
git push origin master
```


GitHub


GitHub Docs
GitHub Terms of Service
GitHub Corporate Terms of Service
GitHub Privacy Statement
GitHub Data Protection Agreement (Non-Enterprise Customers)
Global Privacy Practices
GitHub Insights and data protection for your organization
GitHub Sponsors Additional Terms
GitHub Additional Product Terms
English
Search topics, products...
GitHub.com / Site policy / GitHub Terms of Service
GitHub Terms of Service
Thank you for using GitHub! We're happy you're here. Please read this Terms of Service agreement carefully before accessing or using GitHub. Because it is such an important contract between us and our users, we have tried to make it as clear as possible. For your convenience, we have presented these terms in a short non-binding summary followed by the full legal terms.
In this article
Summary
The GitHub Terms of Service
A. Definitions
B. Account Terms
C. Acceptable Use
D. User-Generated Content
E. Private Repositories
F. Copyright Infringement and

Companies & Projects Using Git

Google FACEBOOK Microsoft LinkedIn NETFLIX PostgreSQL

GNOME eclipse

Companies & Projects Using Git

Google FACEBOOK Microsoft LinkedIn NETFLIX PostgreSQL

RAILS Qt GNOME eclipse

MOST POPULAR

Free

The basics for individuals and organizations

- Unlimited public/private repositories
- 2,000 Actions minutes/month
  Free for public repositories
- 500MB of Packages storage
  Free for public repositories
- Community support

$0 per month

Create a free organization

Team

Advanced collaboration for individuals and organizations

- ← Everything included in Free, plus...
- Protected branches
- Multiple reviewers in pull requests
- Draft pull requests
- Code owners
- Required reviewers
- Pages and Wikis
- 3,000 Actions minutes/month
  Free for public repositories
- 2GB of Packages storage
  Free for public repositories
- Web-based support

$4 per user/month

Continue with Team

Enterprise

Security, compliance, and flexible deployment

- ← Everything included in Team, plus...
- Automatic security and version updates
- SAML single sign-on
- Advanced auditing
- GitHub Connect
- 50,000 Actions minutes/month
  Free for public repositories
- 50GB of Packages storage
  Free for public repositories

EXCLUSIVE ADD-ONS

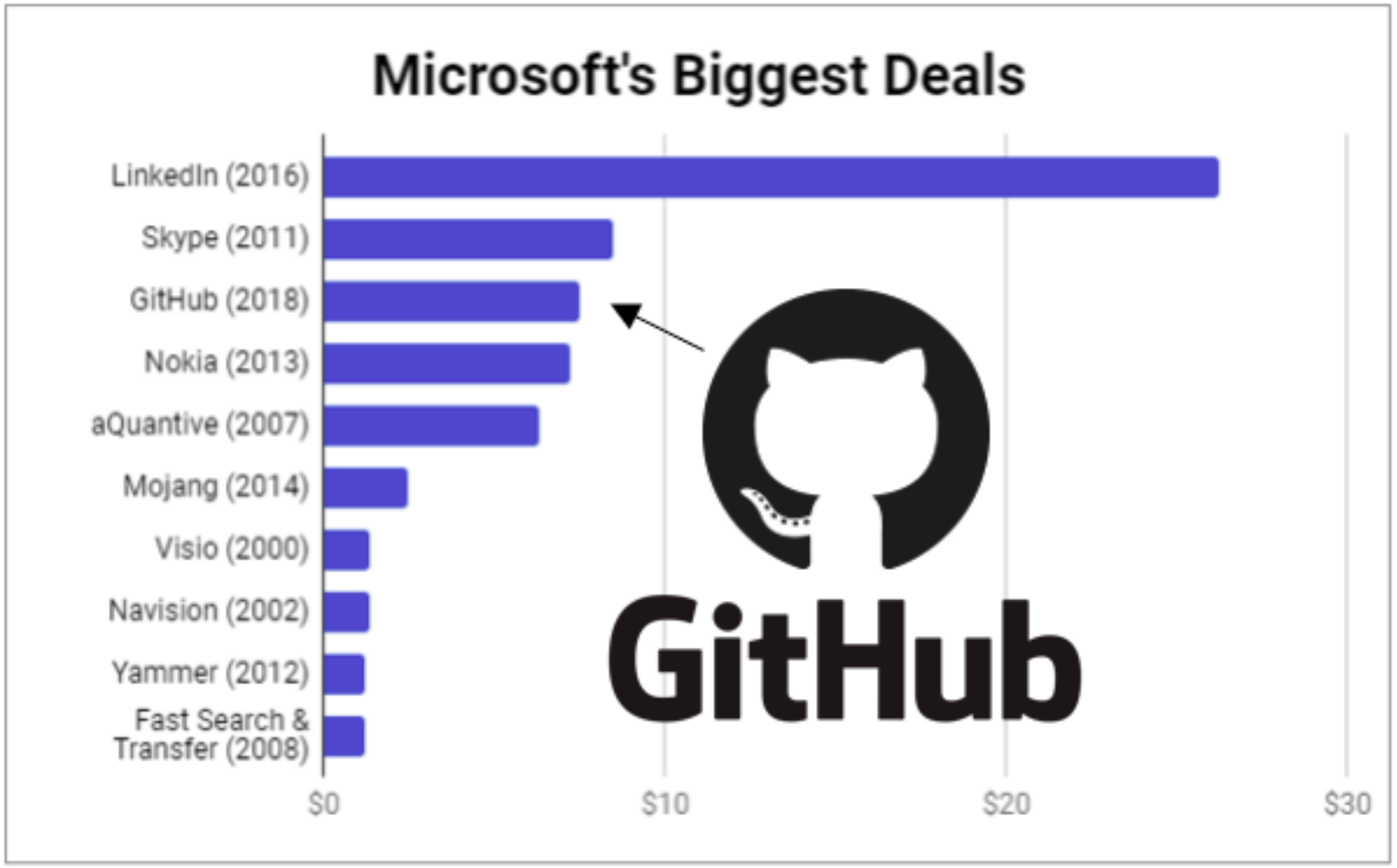
- Token, secret, and code scanning
- Premium support

$21 per user/month

Contact Sales

Microsoft acquires GitHub in 2018 for $7.5 billion

Microsoft acquires GitHub in 2018 for $7.5 billion



Microsoft acquires GitHub in 2018 for $7.5 billion

Tech

GitHub Users Want to Sue Microsoft For Training an AI Tool With Their Code

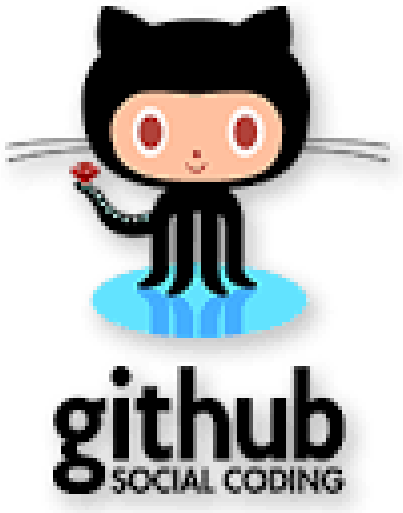
By Janus Rose October 18, 2022, 2:02pm

Share:

https://www.vice.com/en/article/github-users-want-to-sue-microsoft-for-training-an-ai-tool-with-their-code/

So where to put bionformatics software project?

So where to put bionformatics software project?



So where to put bionformatics software project?

Alternatives

ATLASSIAN
Bitbucket

python™
Package
Index

SOURCEFORGE

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

colab

Some of SourceForge's monetization practices have been met with criticism by developers and end users

DevShare adware

Project hijackings and bundled malware

Country restrictions

Some of SourceForge's monetization practices have been met with criticism by developers and end users

DevShare adware

Project hijackings and bundled malware

Country restrictions

https://colab.research.google.com

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- **Zero configuration required**
- **Access to GPUs free of charge**
- **Easy sharing**

https://colab.research.google.com

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- **Zero configuration required**
- **Access to GPUs free of charge**
- **Easy sharing**

alphafold2

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

So where to put bionformatics software project?

Alternatives

ATLASSIAN
Bitbucket

python™
Package
Index

SOURCEFORGE

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

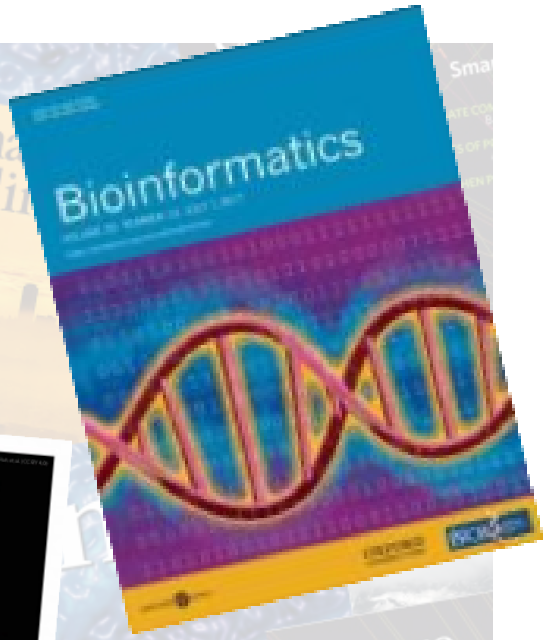
colab

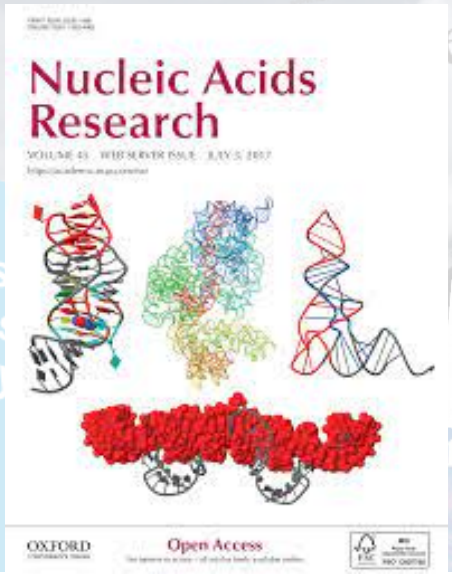
So where to put bioinformatics software project?

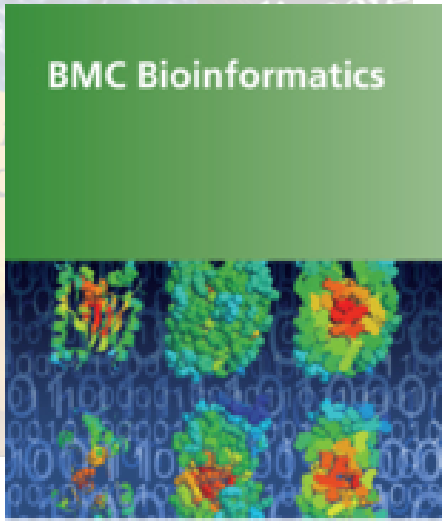
How to announce bioinformatics software project to the world?

How to announce bioinformatics software project to the world?

How to announce bioinformatics software project to the world?







How to announce bioinformatics software project to the world?

How to announce bioinformatics software project to the world?

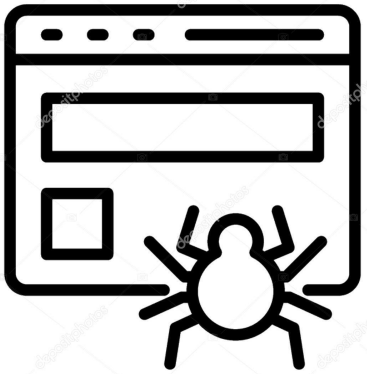
www.biorxiv.org

Verifying you are human. This may take a few seconds.

Verifying...

CLOUDFLARE

Privacy • Terms



Web Crawler

www.biorxiv.org needs to review the security of your connection before proceeding.

How to announce bioinformatics software project to the world?

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

arXiv.org

PeerJ

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

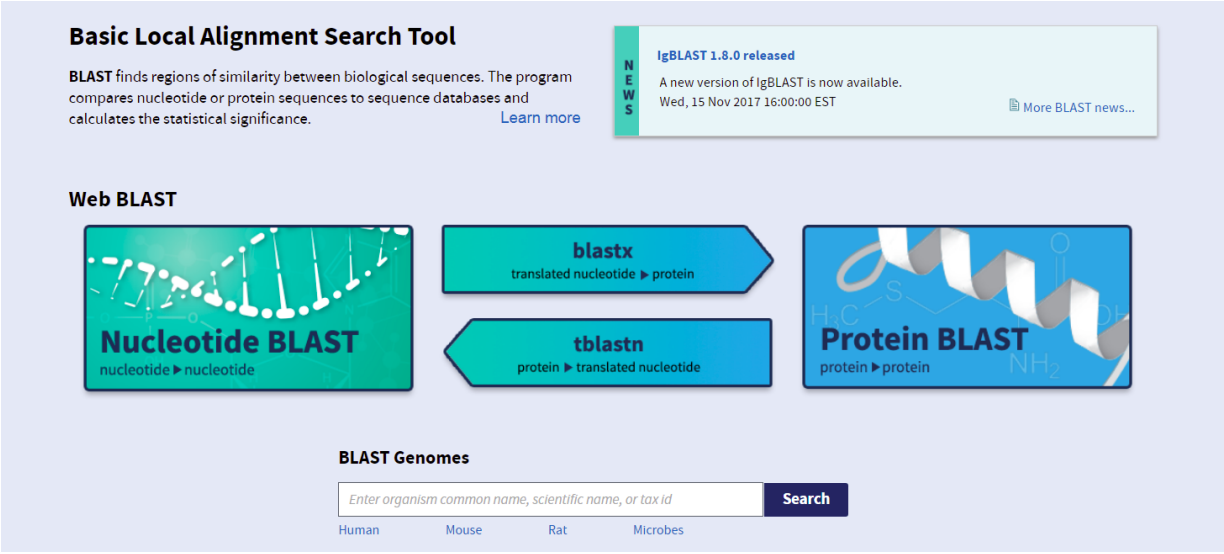
ChemRxiv™

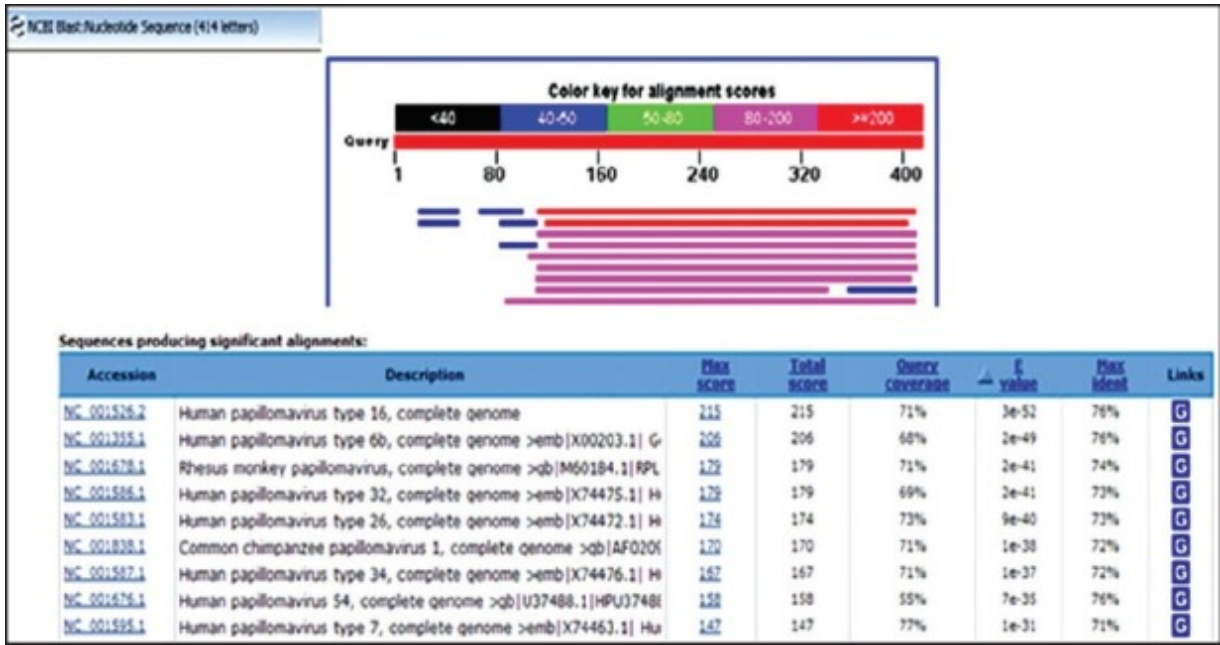
SSRN
tomorrow's research today

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

SOC ARXIV

How to announce bioinformatics software project to the world?





How to announce bioinformatics software project to the world?

J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]Department of Computer Science
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]Department of Computer Science
University of Arizona, Tucson, AZ 85721, U.S.A.

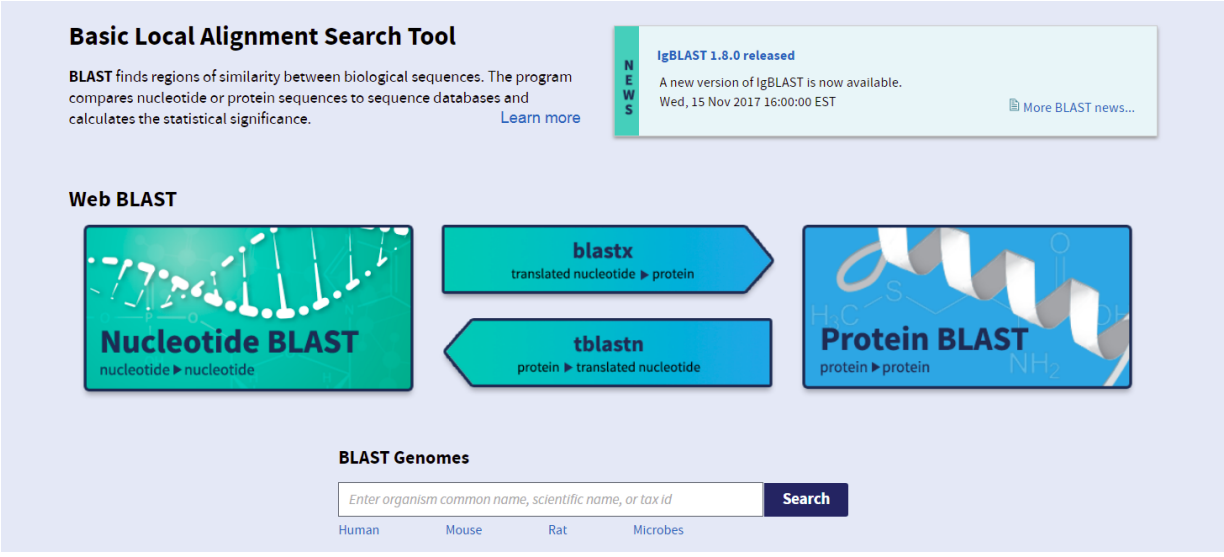
(Received 26 February 1990, accepted 15 May 1990)

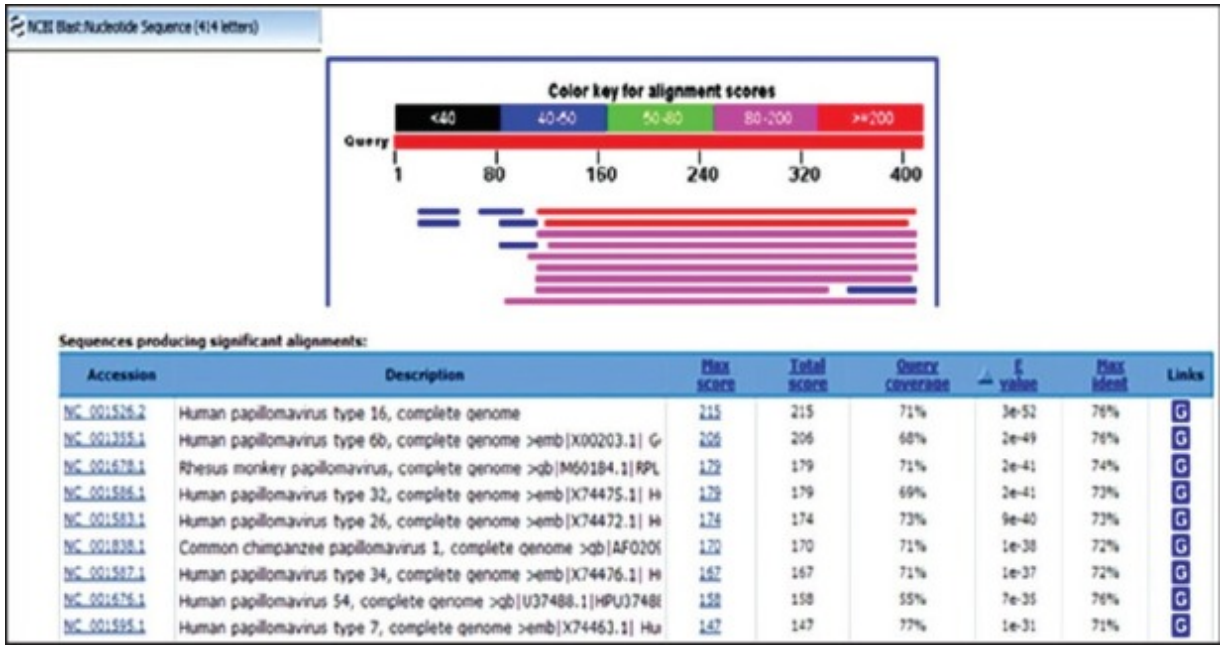
A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

1. Introduction

The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene. As the DNA and amino acid sequence databases continue to grow in size they become increasingly useful in the analysis of newly sequenced genes and optimal, based on the given scores. Because of their computational requirements, dynamic programming algorithms are impractical for searching large databases without the use of a supercomputer (Gotoh & Tagashira, 1986) or other special purpose hardware (Coulson *et al.*, 1987).

Rapid heuristic algorithms that attempt to approximate the above methods have been deve-





How to announce bioinformatics software project to the world?

J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]Department of Computer Science
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]Department of Computer Science
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

1. Introduction

The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene. As the DNA and amino acid sequence databases continue to grow in size they become increasingly useful in the analysis of newly sequenced genes and optimal, based on the given scores. Because of their computational requirements, dynamic programming algorithms are impractical for searching large databases without the use of a supercomputer (Gotoh & Tagashira, 1986) or other special purpose hardware (Coulson *et al.*, 1987).

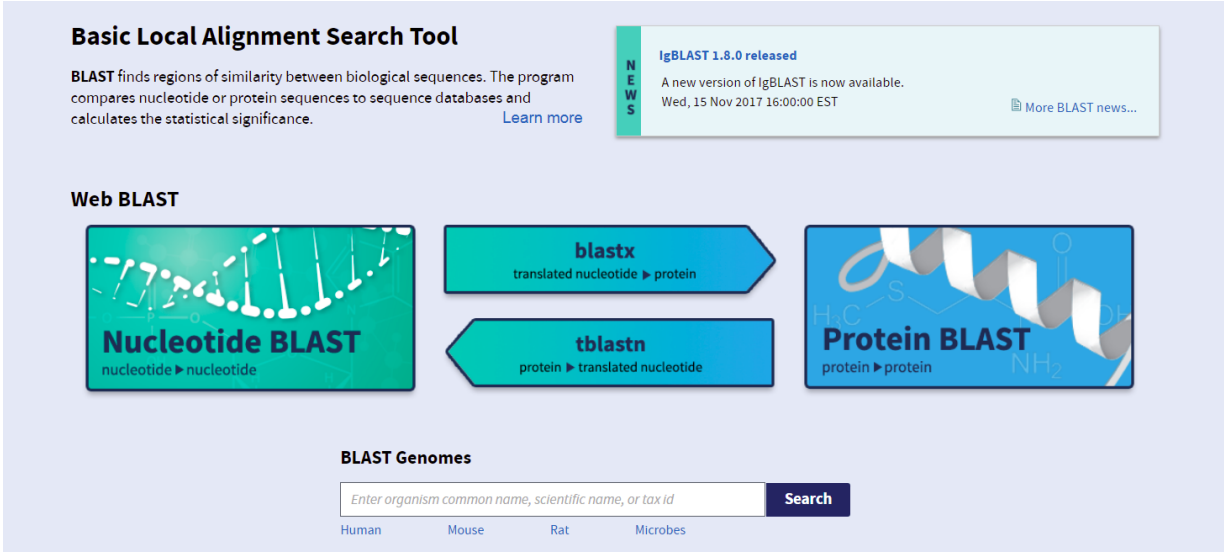
Rapid heuristic algorithms that attempt to approximate the above methods have been deve-

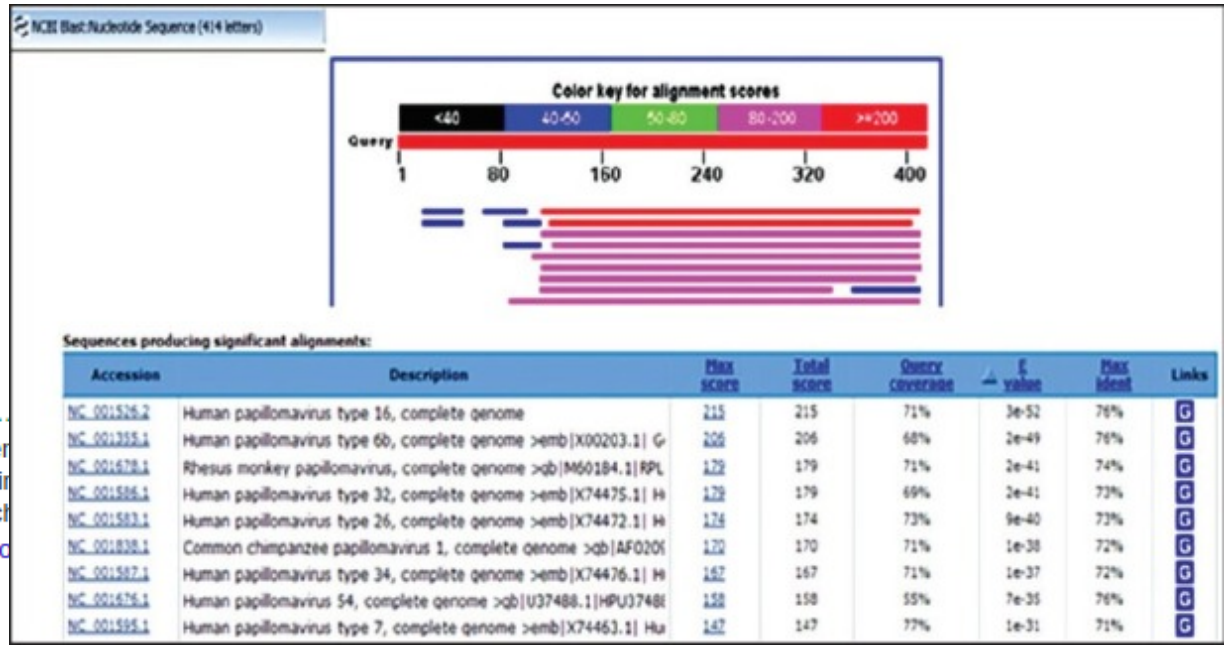
Basic local alignment search tool

SF Altschul, W Gish, W Miller, EW Myers... - Journal of molecular ...

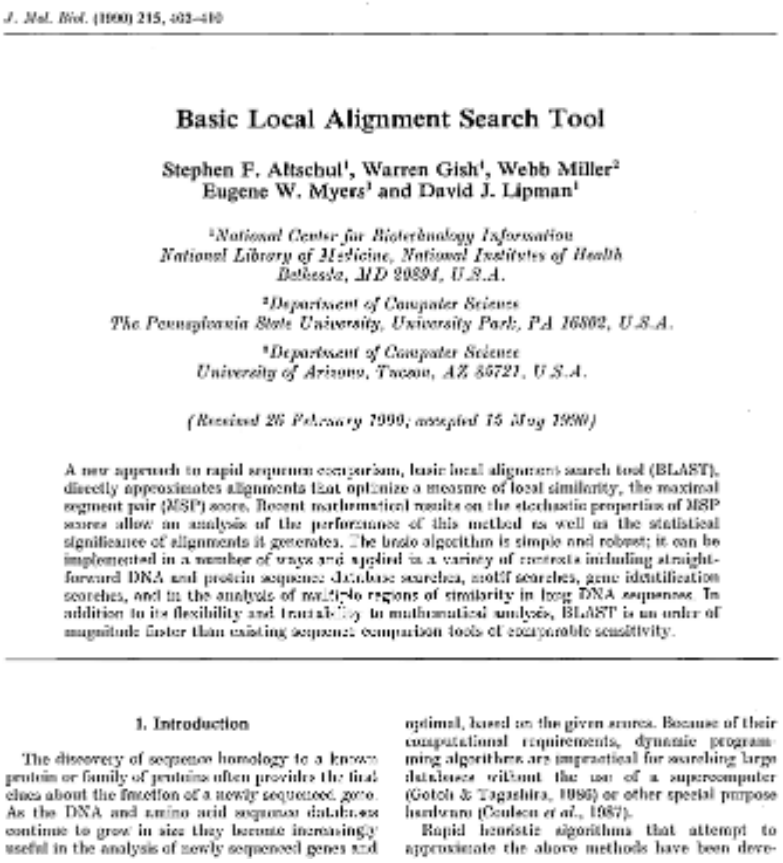
A new approach to rapid sequence comparison, basic local alignmer
directly approximates alignments that optimize a measure of local sir
segment pair (MSP) score. Recent mathematical results on the stoch

☆ 99 Cited by 90873 Related articles All 84 versions Web c





How to announce bioinformatics software project to the world?



J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]Department of Computer Science
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]Department of Computer Science
University of Arizona, Tucson, AZ 85721, U.S.A.

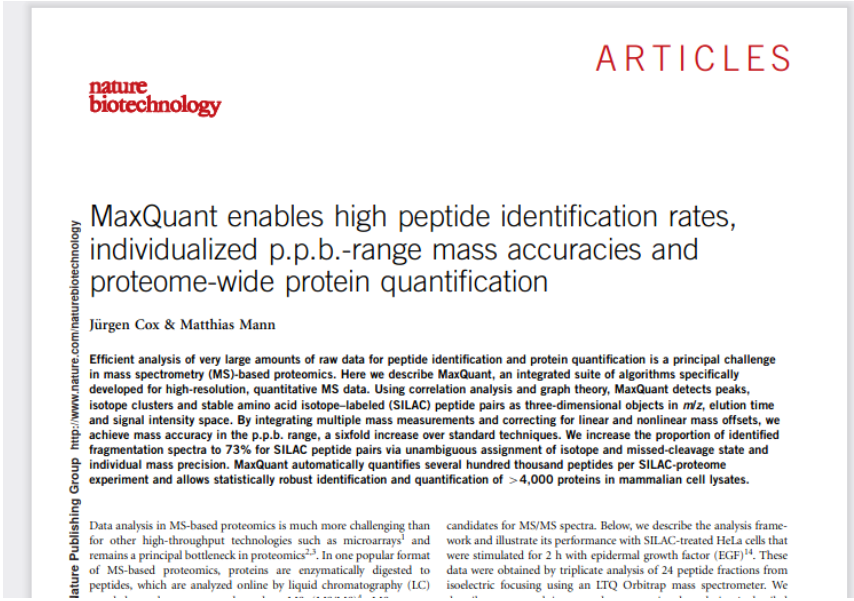
(Received 26 February 1990, accepted 15 May 1990)

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

1. Introduction

The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene. As the DNA and amino acid sequence databases continue to grow in size they become increasingly useful in the analysis of newly sequenced genes and optimal, based on the given scores. Because of their computational requirements, dynamic programming algorithms are impractical for searching large databases without the use of a supercomputer (Gotoh & Tagashira, 1986) or other special purpose hardware (Coulson *et al.*, 1987).

Rapid heuristic algorithms that attempt to approximate the above methods have been deve-



ARTICLES

nature biotechnology

http://www.nature.com/naturebiotechnology Nature Publishing Group

MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification

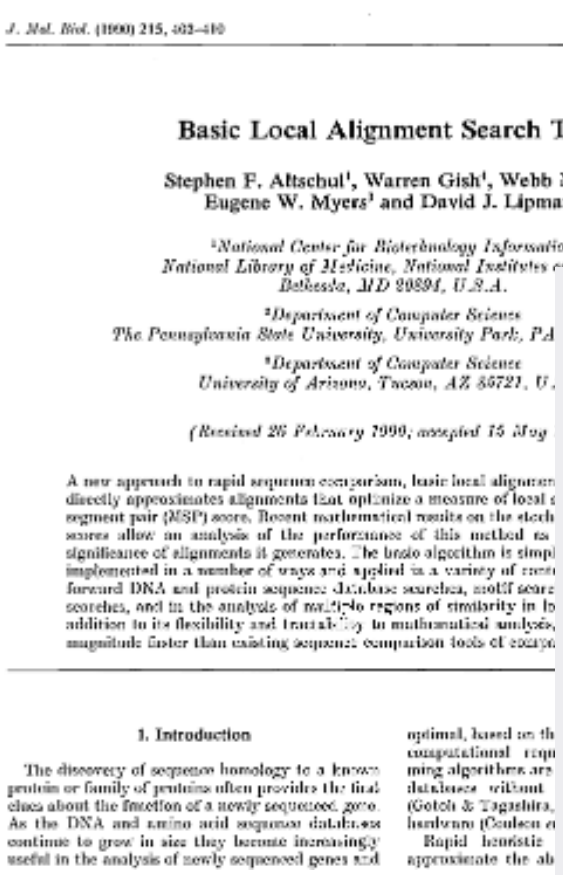
Jürgen Cox & Matthias Mann

Efficient analysis of very large amounts of raw data for peptide identification and protein quantification is a principal challenge in mass spectrometry (MS)-based proteomics. Here we describe MaxQuant, an integrated suite of algorithms specifically developed for high-resolution, quantitative MS data. Using correlation analysis and graph theory, MaxQuant detects peaks, isotope clusters and stable amino acid isotope–labeled (SILAC) peptide pairs as three-dimensional objects in *m/z*, elution time and signal intensity space. By integrating multiple mass measurements and correcting for linear and nonlinear mass offsets, we achieve mass accuracy in the p.p.b. range, a sixfold increase over standard techniques. We increase the proportion of identified fragmentation spectra to 73% for SILAC peptide pairs via unambiguous assignment of isotope and missed-cleavage state and individual mass precision. MaxQuant automatically quantifies several hundred thousand peptides per SILAC-proteome experiment and allows statistically robust identification and quantification of >4,000 proteins in mammalian cell lysates.

Data analysis in MS-based proteomics is much more challenging than for other high-throughput technologies such as microarrays[1] and remains a principal bottleneck in proteomics[2,3]. In one popular format of MS-based proteomics, proteins are enzymatically digested to peptides, which are analyzed online by liquid chromatography (LC)

candidates for MS/MS spectra. Below, we describe the analysis framework and illustrate its performance with SILAC-treated HeLa cells that were stimulated for 2 h with epidermal growth factor (EGF)[14]. These data were obtained by triplicate analysis of 24 peptide fractions from isoelectric focusing using an LTQ Orbitrap mass spectrometer. We

How to announce bionformatics software project to the world?



J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]Department of Computer Science
The Pennsylvania State University, University Park, PA

[3]Department of Computer Science
University of Arizona, Tucson, AZ 85721, U.

(Received 26 February 1990; accepted 15 May

A new approach to rapid sequence comparison, basic local alignment
directly approximates alignments that optimize a measure of local
segment pair (MSP) score. Recent mathematical results on the stoch
scores allow an analysis of the performance of this method as
significance of alignments it generates. The basic algorithm is simp
implemented in a number of ways and applied in a variety of recent
forward DNA and protein sequence database searches, motif searc
searches, and in the analysis of multiple regions of similarity in lo
addition to its flexibility and tractability to mathematical analysis,
magnitude faster than existing sequence comparison tools of accep

1. Introduction

The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene. As the DNA and amino acid sequence databases continue to grow in size they become increasingly useful in the analysis of newly sequenced genes and

optimal, based on th
computational requ
ming algorithms are
databases without
(Gotoh & Tagashira,
hardware (Coulson e
Rapid heuristic
approximate the ab

MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods

Koichiro Tamura,[1,2] Daniel Peterson,[2] Nicholas Peterson,[2] Glen Stecher,[2] Masatoshi Nei,[3] and Sudhir Kumar*[,2,4]

[1]Department of Biological Sciences, Tokyo Metropolitan University, Hachioji, Tokyo, Japan
[2]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University
[3]Department of Biology and the Institute of Molecular Evolutionary Genetics, The Pennsylvania State University
[4]School of Life Sciences, Arizona State University
*Corresponding author: E-mail: s.kumar@asu.edu.
Associate editor: Naoki Takebayashi

Abstract

Comparative analysis of molecular sequence data is essential for reconstructing the evolutionary histories of species and inferring the nature and extent of selective forces shaping the evolution of genes and species. Here, we announce the release of Molecular Evolutionary Genetics Analysis version 5 (MEGA5), which is a user-friendly software for mining online databases, building sequence alignments and phylogenetic trees, and using methods of evolutionary bioinformatics in basic biology, biomedicine, and evolution. The newest addition in MEGA5 is a collection of maximum likelihood (ML) analyses for inferring evolutionary trees, selecting best-fit substitution models (nucleotide or amino acid), inferring ancestral states and sequences (along with probabilities), and estimating evolutionary rates site-by-site. In computer simulation analyses, ML tree inference algorithms in MEGA5 compared favorably with other software packages in terms of computational efficiency and the accuracy of the estimates of phylogenetic trees, substitution parameters, and rate variation among sites. The MEGA user interface has now been enhanced to be activity driven to make it easier for the use of both beginners and experienced scientists. This version of MEGA is intended for the Windows platform, and it has been configured for effective use on Mac OS X and Linux desktops. It is available free of charge from http://www.megasoftware.net.

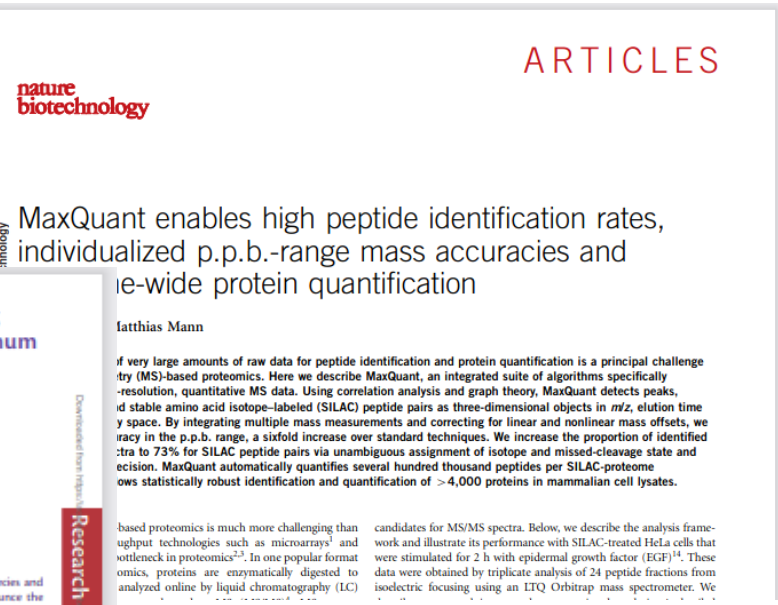
Introduction

The Molecular Evolutionary Genetics Analysis (MEGA) software was developed with the goal of providing a biologist centric, integrated suite of tools for statistical analyses of DNA and protein sequence data from an evolutionary standpoint. Over the years, it has grown to include tools for sequence alignment, phylogenetic tree reconstruction and visualization, testing an array of evolutionary hypotheses, estimating sequence divergences, web-based acquisition of sequence data, and expert systems to generate natural language descriptions of the analysis methods and data chosen by the user (Kumar et al. 1994, 2008; Kumar and Dudley 2007). With the fifth major release, the collection of analysis tools in MEGA has now broadened to include the maximum likelihood (ML) methods for molecular evolutionary analysis. Table 1 contains a summary of all statistical methods and models in MEGA5, with new features marked with an asterisk (*). In the following, we provide a brief description of methodological advancements, along with relevant research results, and technical enhancements in MEGA5.

Model Selection for Nucleotide and Amino Acid sequences

MEGA5 now contains facilities to evaluate the fit of major models of nucleotide and amino acid substitutions, which are frequently desired by researchers (Posada and Crandall 1998; Nei and Kumar 2000; Yang 2006) (fig. 1A). For nucleotide substitutions, the GTR and five nested models are available, whereas six models with and without empirical frequencies (+F) have been programmed for the amino acid substitutions (Table 1). MEGA5 provides the goodness-of-fit (see below) of the substitution models with and without assuming the existence of evolutionary rate variation among sites, which is modeled by a discrete Gamma distribution (+G) (Yang 1994) and/or an allowance for the presence of invariant sites (+I) (Fitch and Margoliash 1967; Fitch 1986; Shoemaker and Fitch 1989). This results in an evaluation of 24 and 48 models for nucleotide and amino acid substitutions, respectively. For each of these models, MEGA5 provides the estimated values of shape parameter of the Gamma distribution (α), the proportion of invariant sites, and the substitution rates between bases or residues, as applicable. Depending on the model, the assumed or observed values of the base or amino acid frequencies used in the analysis are also provided. This information enables researchers to quickly examine the robustness of the estimates of evolutionary parameters under different models of substitutions and assumptions about the distribution of evolutionary rates among sites (fig. 1C). The goodness-of-fit of each model to the data is measured by the Bayesian information criterion (BIC, Schwarz 1978) and corrected

Research resource



ARTICLES

nature biotechnology

MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and ...e-wide protein quantification

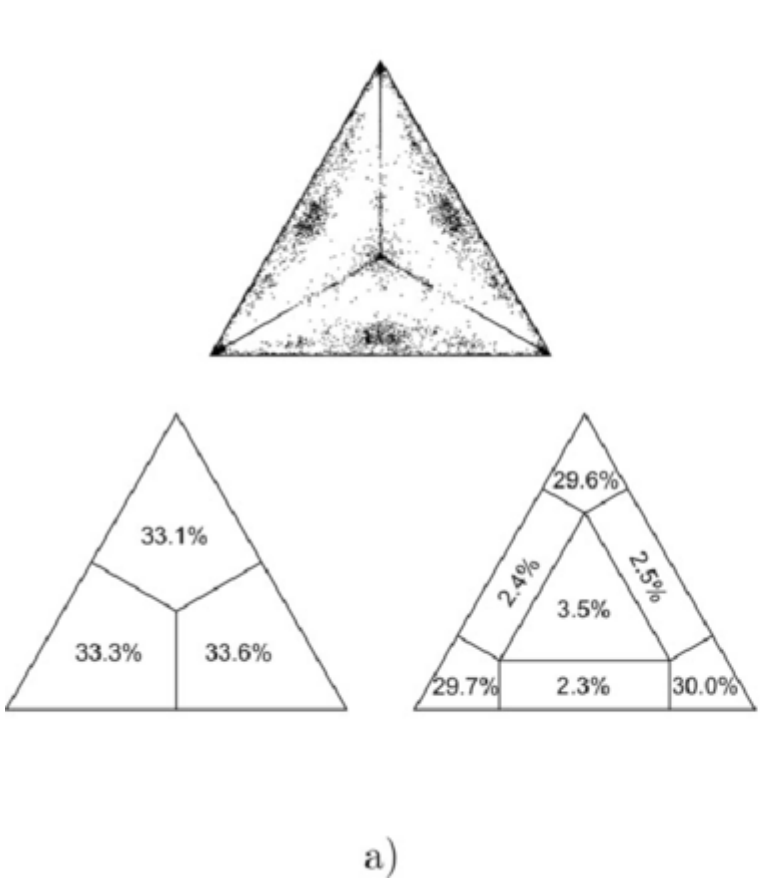
...latthias Mann

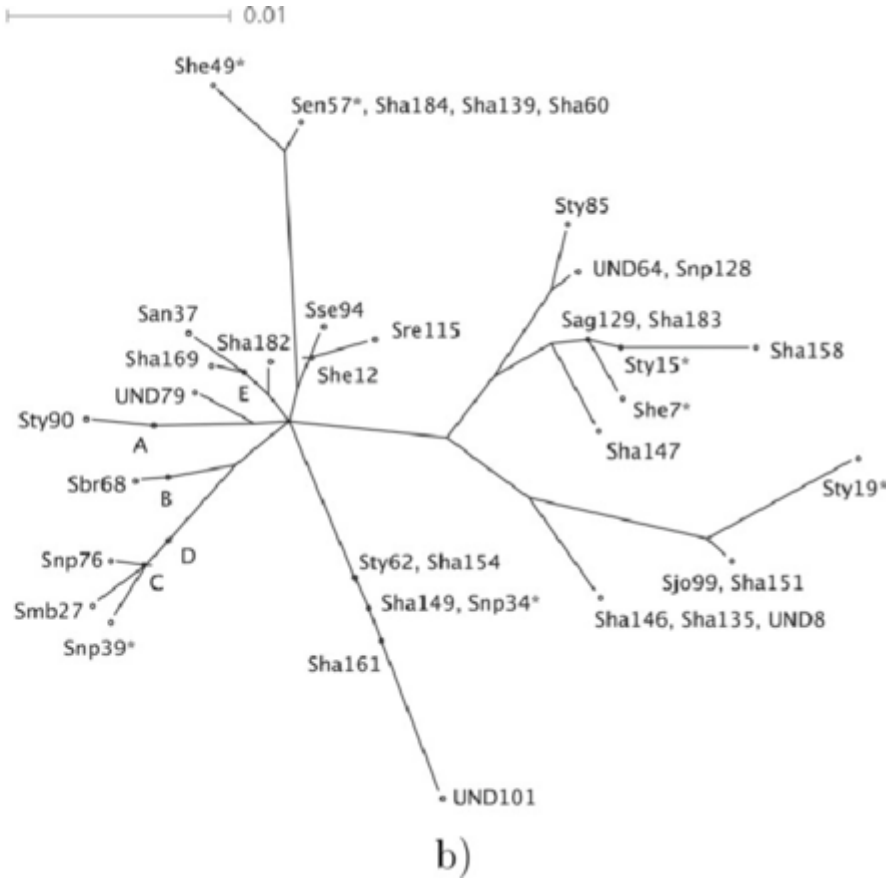
...f very large amounts of raw data for peptide identification and protein quantification is a principal challenge ...try (MS)-based proteomics. Here we describe MaxQuant, an integrated suite of algorithms specifically ...-resolution, quantitative MS data. Using correlation analysis and graph theory, MaxQuant detects peaks, ...d stable amino acid isotope–labeled (SILAC) peptide pairs as three-dimensional objects in *m/z*, elution time ...y space. By integrating multiple mass measurements and correcting for linear and nonlinear mass offsets, we ...racy in the p.p.b. range, a sixfold increase over standard techniques. We increase the proportion of identified ...tra to 73% for SILAC peptide pairs via unambiguous assignment of isotope and missed-cleavage state and ...ecision. MaxQuant automatically quantifies several hundred thousand peptides per SILAC-proteome ...ows statistically robust identification and quantification of >4,000 proteins in mammalian cell lysates.

...based proteomics is much more challenging than ...ughput technologies such as microarrays[1] and ...ottleneck in proteomics[2,3]. In one popular format ...omics, proteins are enzymatically digested to ...analyzed online by liquid chromatography (LC)

candidates for MS/MS spectra. Below, we describe the analysis framework and illustrate its performance with SILAC-treated HeLa cells that were stimulated for 2 h with epidermal growth factor (EGF)[14]. These data were obtained by triplicate analysis of 24 peptide fractions from isoelectric focusing using an LTQ Orbitrap mass spectrometer. We

How to announce bioinformatics software project to the world?





TREEFINDER

BMC Ecology and Evolution

to

Retraction Note | Open Access | Published: 05 November 2015

Retraction Note: TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics

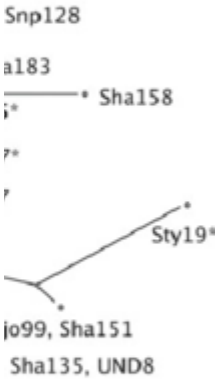
Gangolf Jobb ✉, Arndt von Haeseler & Korbinian Strimmer

BMC Evolutionary Biology **15**, Article number: 243 (2015) | Cite this article

The Original Article was published on 28 June 2004

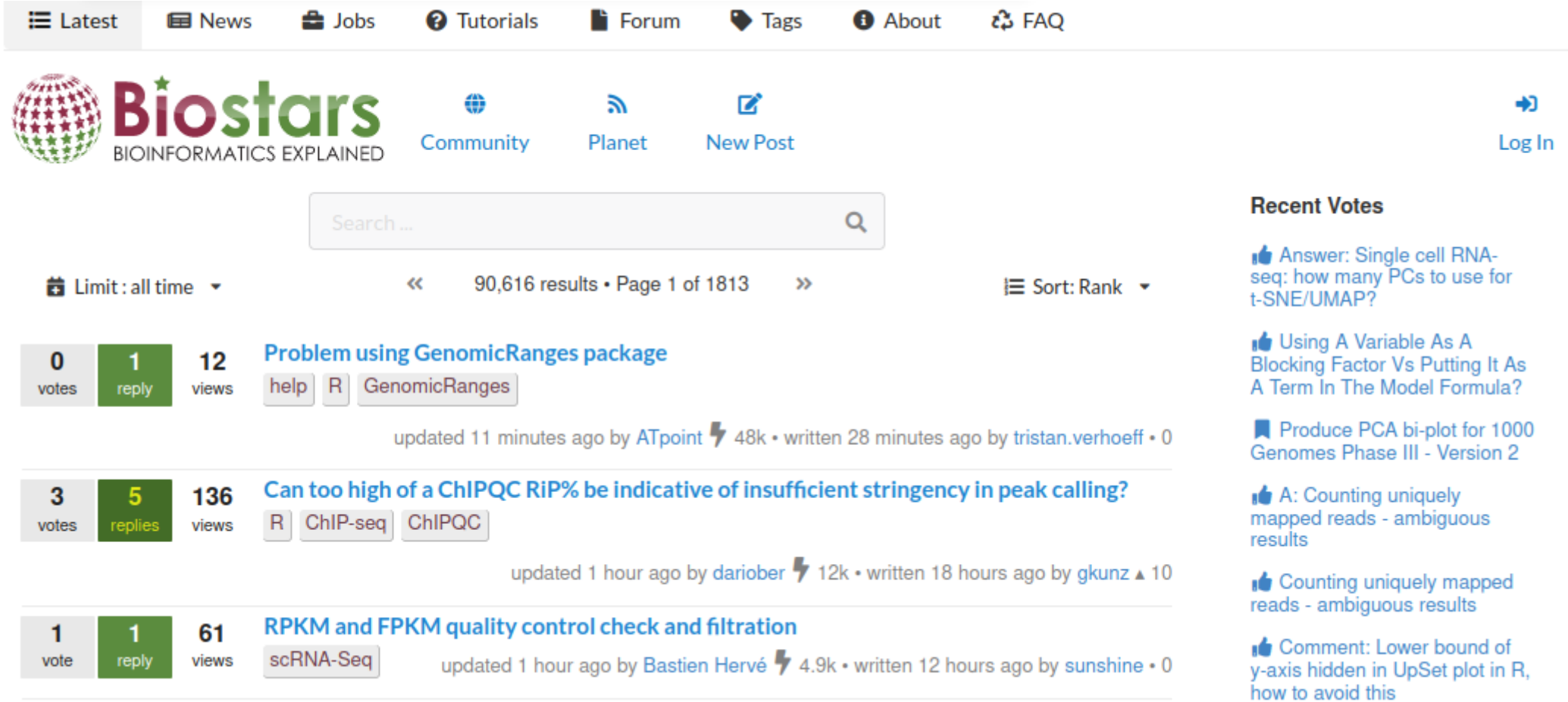
Retraction

The editors of *BMC Evolutionary Biology* retract this article [1] due to the decision by the corresponding author, Gangolf Jobb, to change the license to the software described in the article. The software is no longer available to all scientists wishing to use it in certain territories. This breaches the journal's editorial policy on software availability [2] which has been in effect since the time of publication. The other authors of the article, Arndt von Haeseler and Korbinian Strimmer, have no control over the licensing of the software and support the retraction of this article.



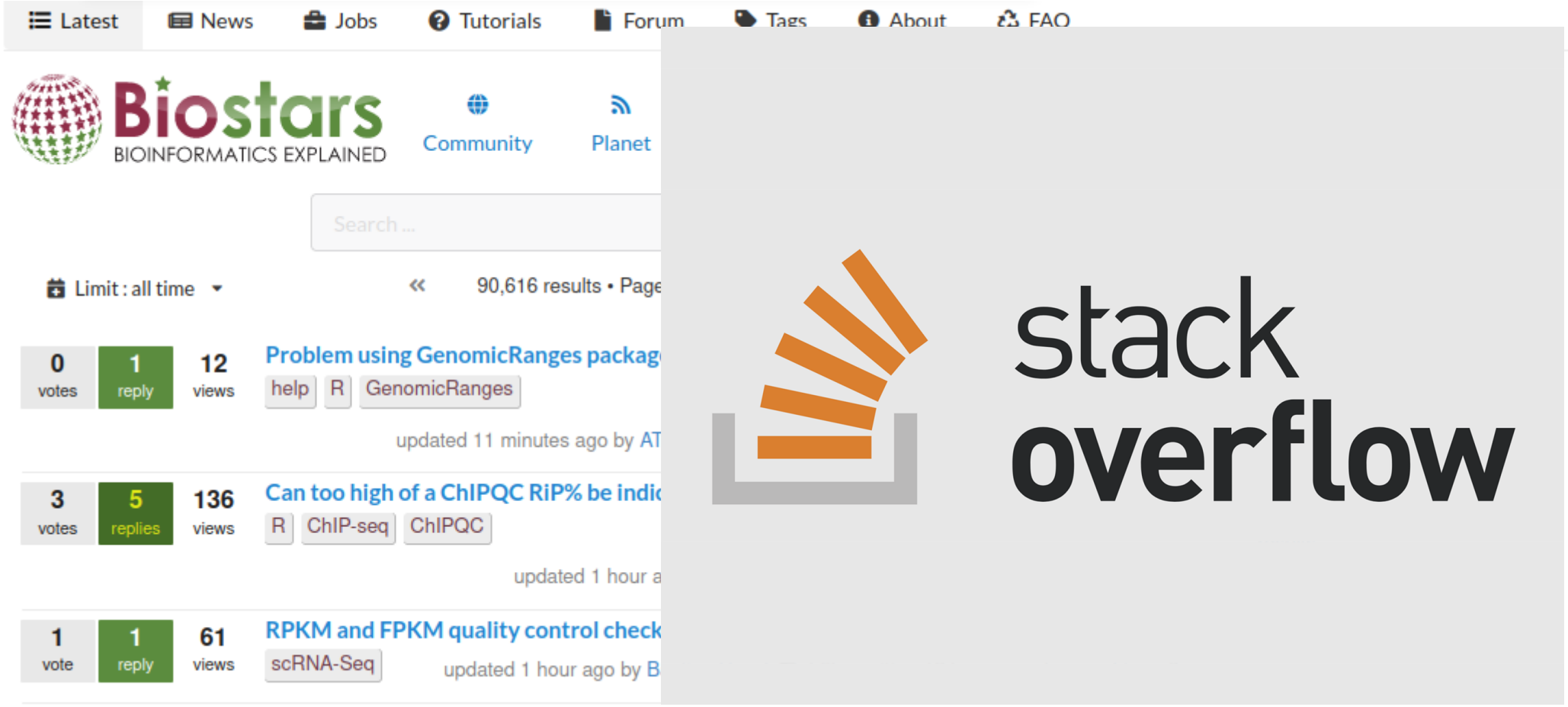
MUST KNOWN FOR BIOINFORMATICIAN

https://www.biostars.org/



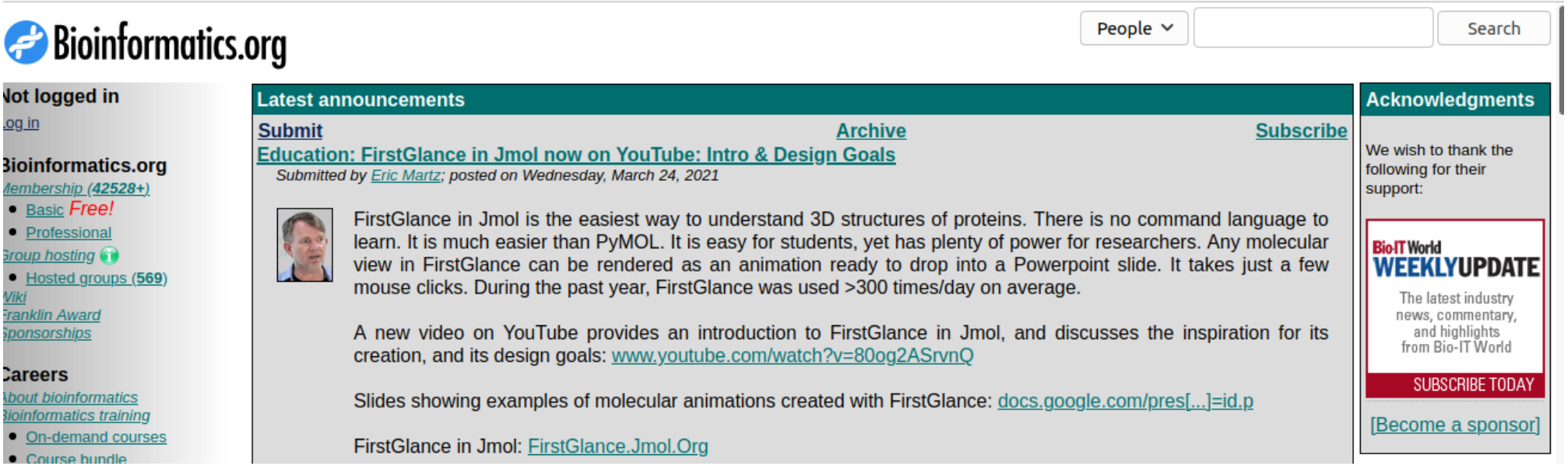
MUST KNOWN FOR BIOINFORMATICIAN

https://www.biostars.org/



MUST KNOWN FOR BIOINFORMATICIAN

https://www.bioinformatics.org/



MUST KNOWN FOR BIOINFORMATICIAN

https://fairsharing.org/

FAIRsharing.org
standards, databases, policies

https://bio.tools/

elixir bio.tools

CREDITS:

The presentation is partly based on

https://www.freecodecamp.org/news/the-beginners-guide-to-git-github/

https://www.slideshare.net/LarsYde1/version-control-issue-tracking-and-communication

https://guides.github.com/activities/hello-world/

https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners

https://www.edureka.co/blog/how-to-use-github/

https://www.freecodecamp.org/news/the-beginners-guide-to-git-github/

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

lukaskoz@mimuw.edu.pl

Zhu *et al. Journal of Hematology & Oncology* (2024) 17:27
https://doi.org/10.1186/s13045-024-01543-8

Journal of Hematology & Oncology

CORRESPONDENCE Open Access

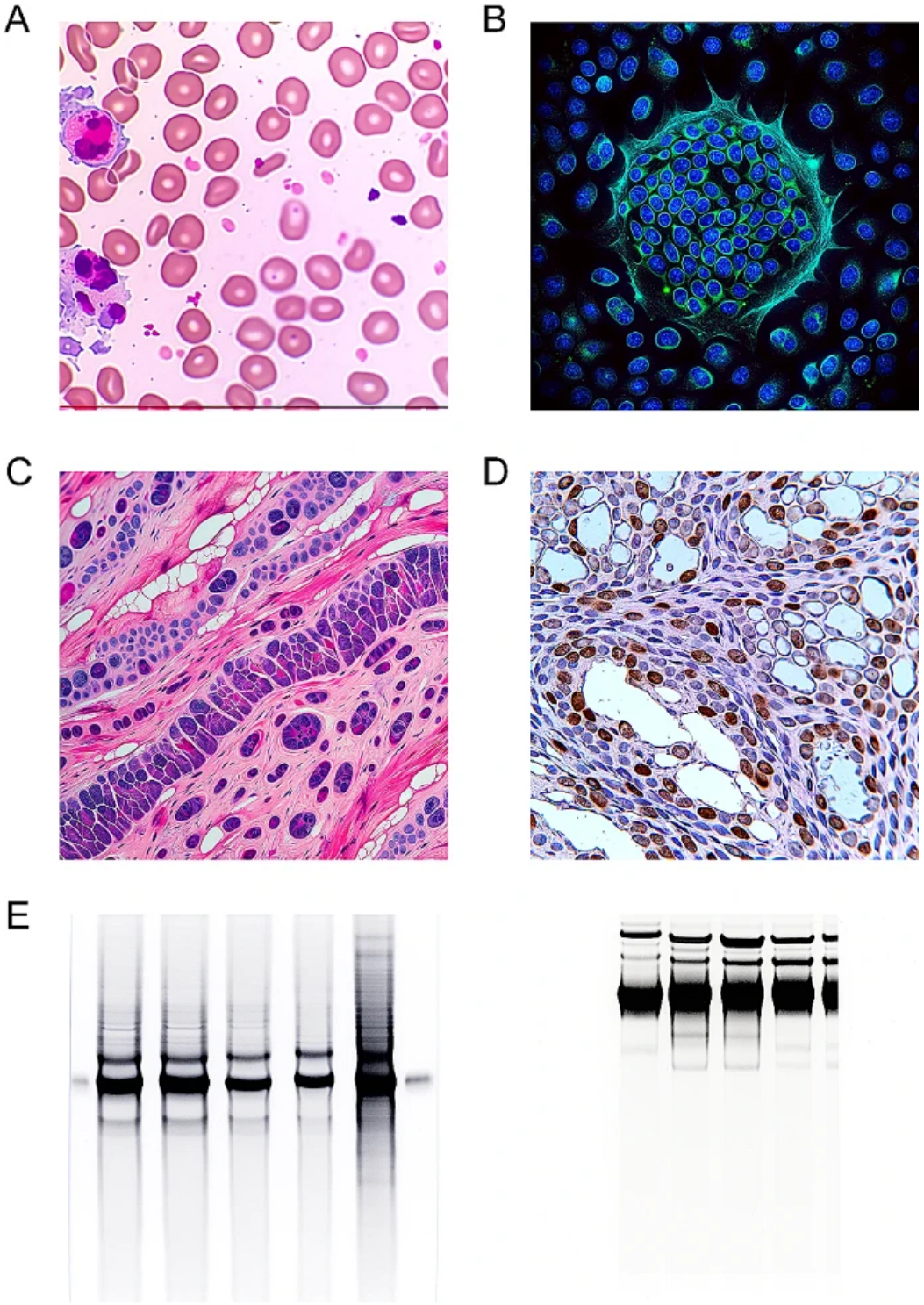
ChatGPT's ability to generate realistic experimental images poses a new challenge to academic integrity

Lingxuan Zhu[1†], Yancheng Lai[1†], Weiming Mou[2†], Haoran Zhang[1], Anqi Lin[1], Chang Qi[3], Tao Yang[4], Liling Xu[1], Jian Zhang[1*] and Peng Luo[1*]

Abstract

The rapid advancements in large language models (LLMs) such as ChatGPT have raised concerns about their potential impact on academic integrity. While initial concerns focused on ChatGPT's writing capabilities, recent updates have integrated DALL-E 3's image generation features, extending the risks to visual evidence in biomedical research. Our tests revealed ChatGPT's nearly barrier-free image generation feature can be used to generate experimental result images, such as blood smears, Western Blot, immunofluorescence and so on. Although the current ability of ChatGPT to generate experimental images is limited, the risk of misuse is evident. This

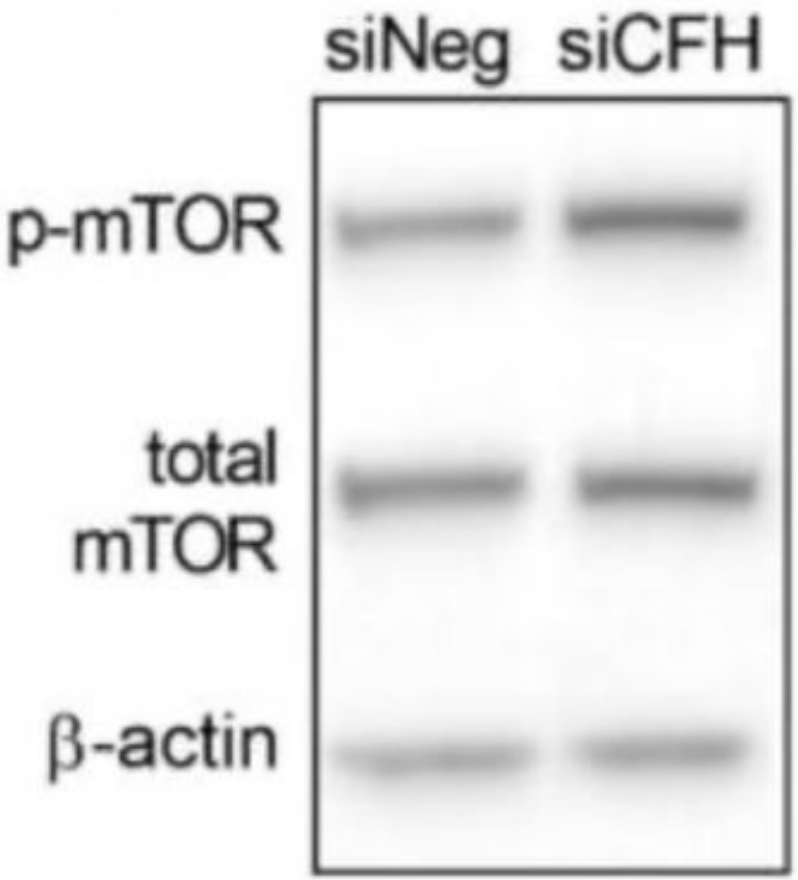
https://doi.org/10.1186/s13045-024-01543-8



Realistic experimental images generated using ChatGPT. (A) blood smears. (B) immunofluorescence staining. (C) hematoxylin and eosin (H&E) staining. (D) immunohistochemistry. (E) western blot images

https://doi.org/10.1186/s13045-024-01543-8

Original



Gen AI