Advanced development of Lancet, an emerging tool for complex variant calling in cancer genomics

Giuseppe Narzisi, PhD

Lead Bioinformatics Scientist

ITCR PI Teleconference
Dec 3 2021

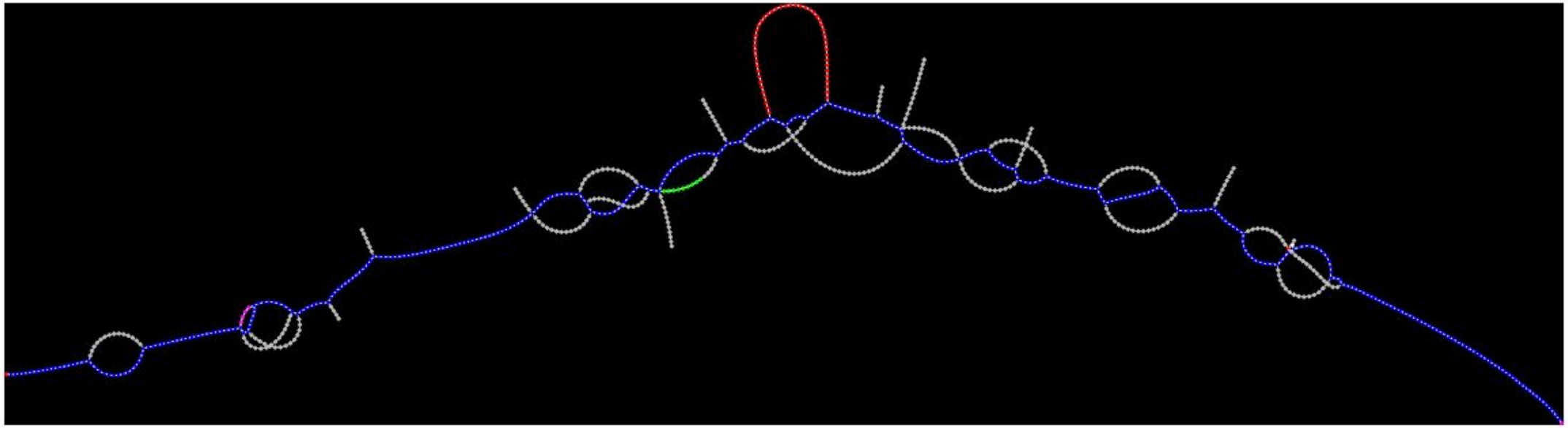
@gnarzisi

gnarzisi@nygenome.org

Lancet: somatic variant calling using colored de Bruijn graphs

- Joint assembly of tumor and normal data
- High accuracy and sensitivity at low VAF
- Outperform state-of-the-art methods in the detection of ‘twilight zone’ indels (30-250 bp)
- Automatic tuning of k-mer and graph structure
- Graph rendering and visualization

red = tumor, green = normal, blue = shared, grey = low coverage & sequencing errors

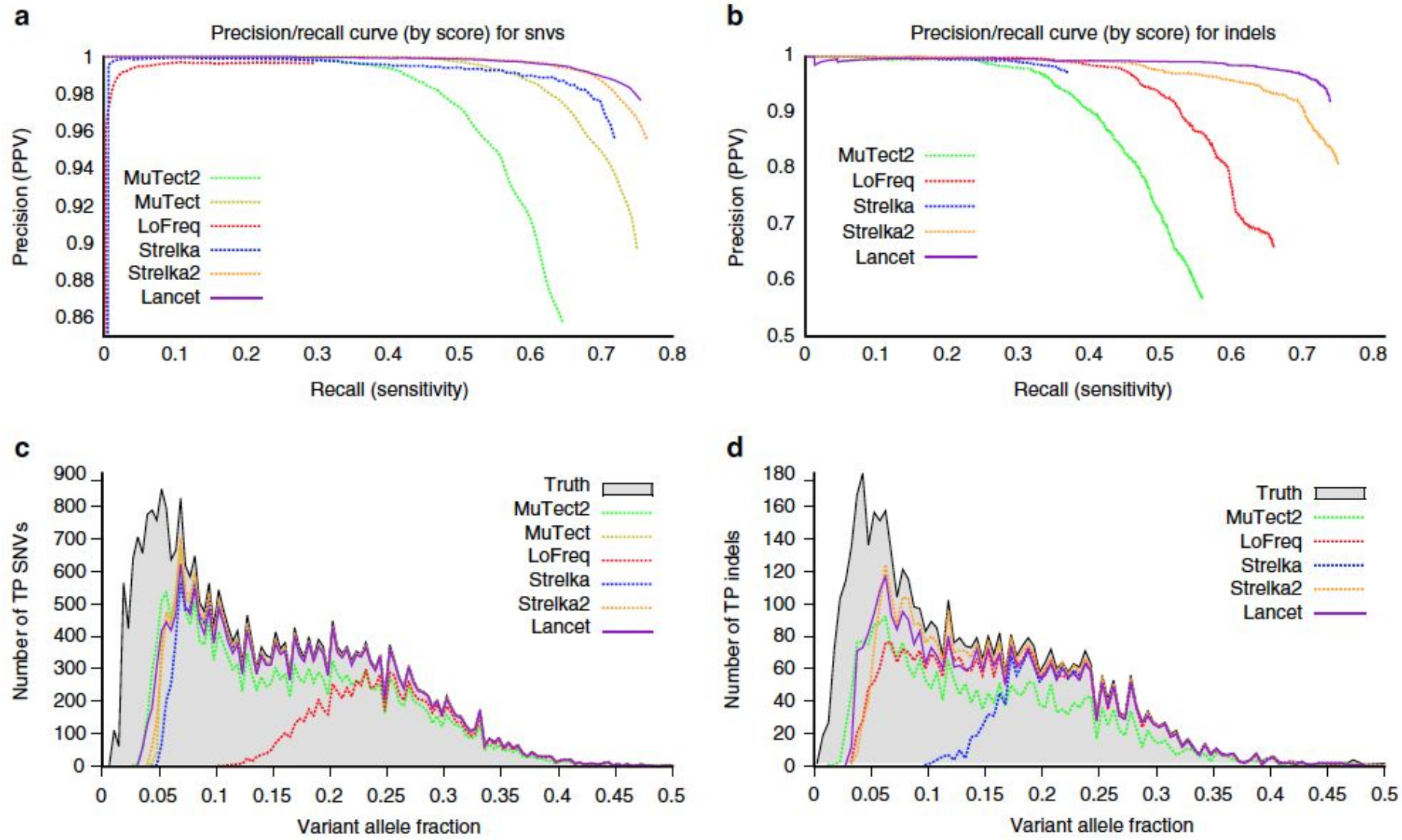


Narzisi G, et al. *Genome-wide somatic variant calling using localized colored de Bruijn graphs.* **Communications Biology**, volume 1, Article number: 20 (2018)

Lancet history

- First released in late-2017.
- Quickly become the method of choice for somatic variant calling via local assembly.
- Funded by ITCR in 2018 under the NCI R21 award 1R21CA220411-01A1.
- Cited in > 40 high-impact papers studying diverse cancer types: melanocytic naevi, polymorphous adenocarcinoma, pancreatic cancers, colorectal cancer, hyalinizing trabecular tumors of the thyroid, triple-negative breast cancer, etc.
- Lancet is a critical component of the cancer pipeline at the New York Genome Center [Arora et al. (2019) Scientific reports, 9(1), 19123]

Somatic mutations performance comparison



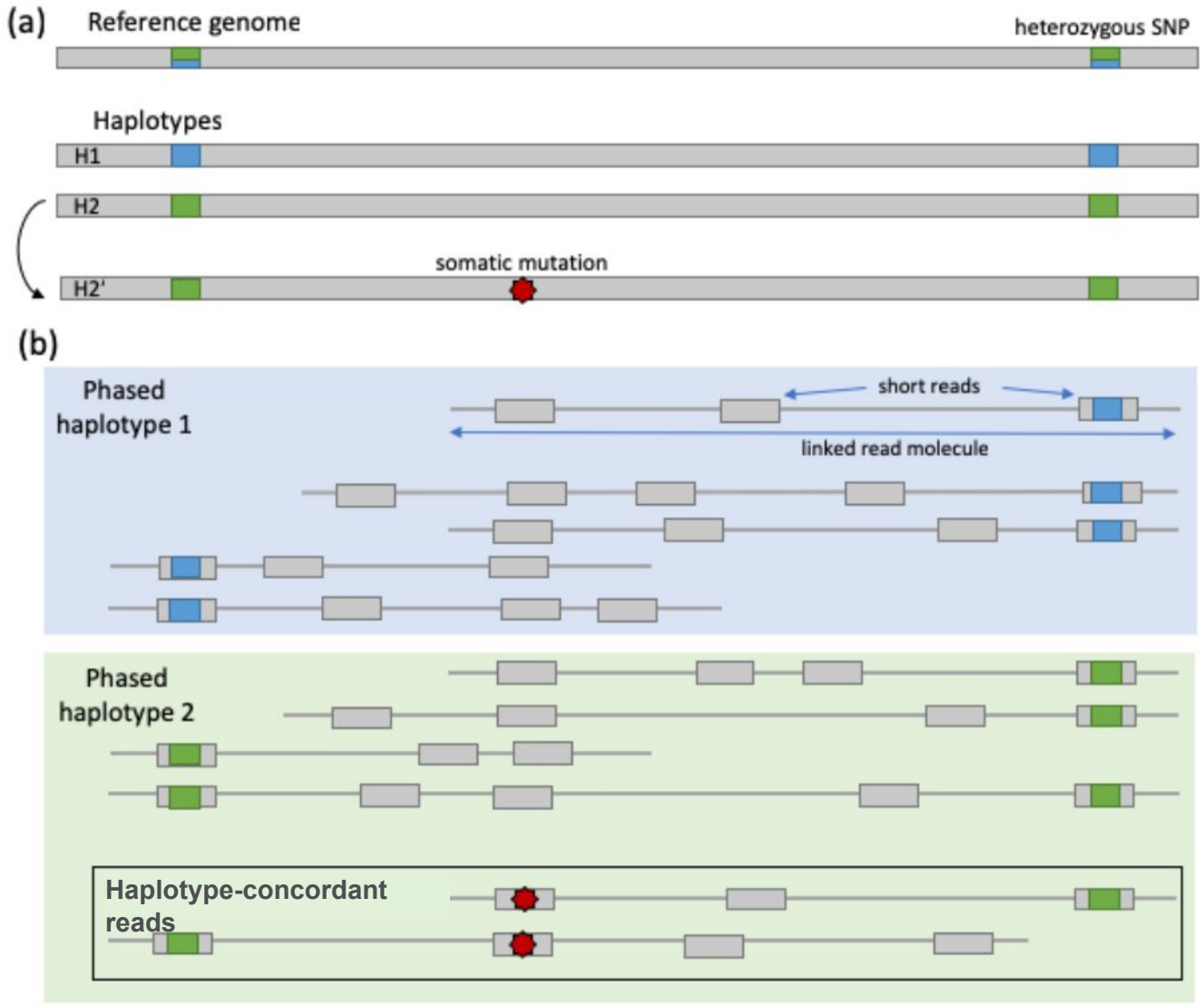
Narzisi G, et al. *Genome-wide somatic variant calling using localized colored de Bruijn graphs.* **Communications Biology**, volume 1, Article number: 20 (2018)

Linked-reads to distinguish real somatic mutations from sequencing errors

a. If a real mosaic variant (red star) arises near a heterozygous mutation it will always be found in conjunction with **only one** of the two alleles (green) and will never appear on reads with the other allele (blue).

b. This generates three haplotypes in bulk sequencing

Conclusion: erroneous somatic calls will manifest themselves as mutations supported by reads on both haplotypes (e.g., phased with mutations occurring on different haplotypes).



From Darby et al. bioRxiv 560532; doi: https://doi.org/10.1101/560532

Musunuri et al., Somatic variant analysis of linked-reads sequencing data with Lancet, **Bioinformatics**, Volume 37, Issue 13, 2021.

Adapted from Figure 3 of Yanmei Dou, Heather D. Gold, Lovelace J. Luquette, and Peter J. Park. "Detecting Somatic Mutations in Normal Cells". In: **Trends in Genetics** 34.7 (2018), pp. 545–557.

Github repo and documentation

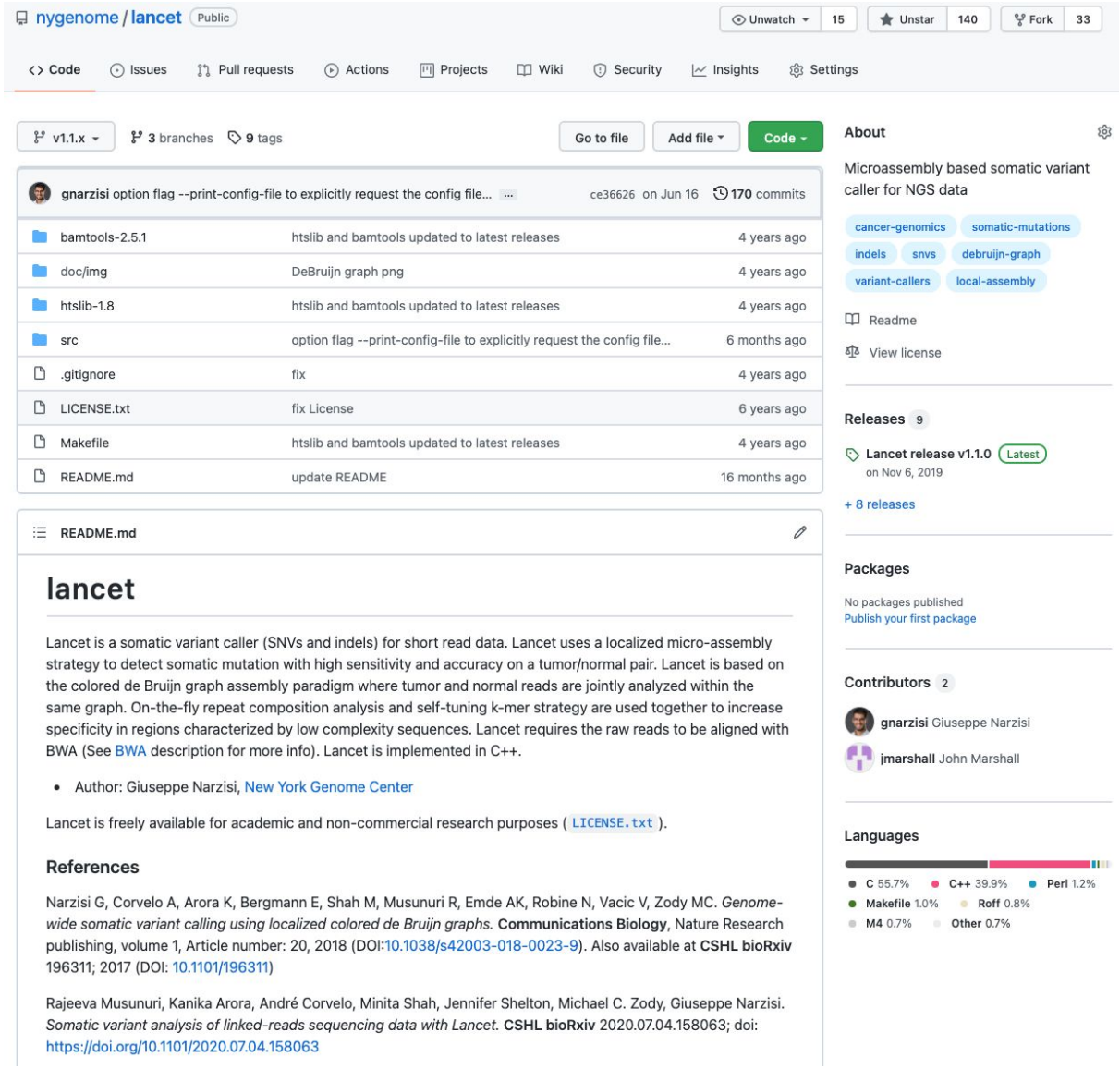
- Source code freely available for academic and non-commercial research purposes via NYGC github: https://github.com/nygenome/lancet

- 100% C/C++ code (no dependencies) with native pthreads parallelization

- Interactive user interface similar to other bioinformatics utilities (e.g., samtools, bamtools, bedtools, etc.).

- APIs and libraries (included):
  - BamTools: https://github.com/pezmaster31/bamtools
  - HTSlib: http://www.htslib.org/

- Compilation:
  1. git clone git://github.com/nygenome/lancet.git
  2. cd lancet
  3. make



U01 Specific Aims

- **Aim 1**. Increase computational performance and facilitate user adoption and third-party development.

- **Aim 2**. Add new features and enhancements to improve variant detection, phasing, and data visualization.

- **Aim 3**. Enable joint assembly and analysis of longitudinal and cohort data.

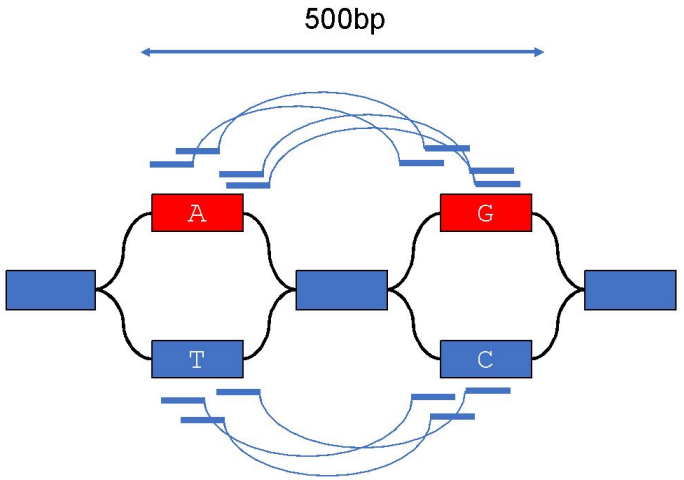
Aim 1: *Software engineering*

Tools	Runtime (core hrs)	Peak Memory (GB)	CPU Utilization (%)
Lancet v1.1.0	26.63	6.41	92.67
Lancet v2 prototype	5.49	5.84	98.45
Octopus v0.6.3-beta	35.03	13.01	65.84
MuTect2 v4.0.5.1	5.10	3.89	33.47
Strelka2 v2.9.3	0.28	0.09	9.17

Computational performance comparison on chr22 of the Virtual Tumor.

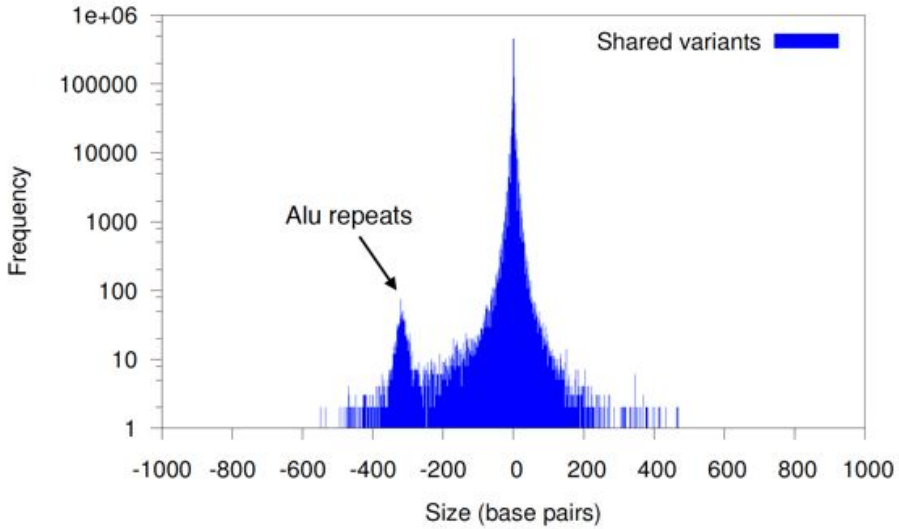
1. Re-factor the source code using modern C++14 features;
2. Plan to integrate state-of-the-art high-performance libraries for accelerated local alignment (ksw2, edlib, parasail) and efficient hash table to store the graph (abseil);
3. Design a developer tool kit to facilitate new feature development and integration in future bioinformatics tools.

Aim 2: *Enhance the core algorithms*



H1: …ACGAAATAACACGTACATTCAAGTCGTATT…
H2: …ACGATATTACACGCACATTCAACTCGTATT…

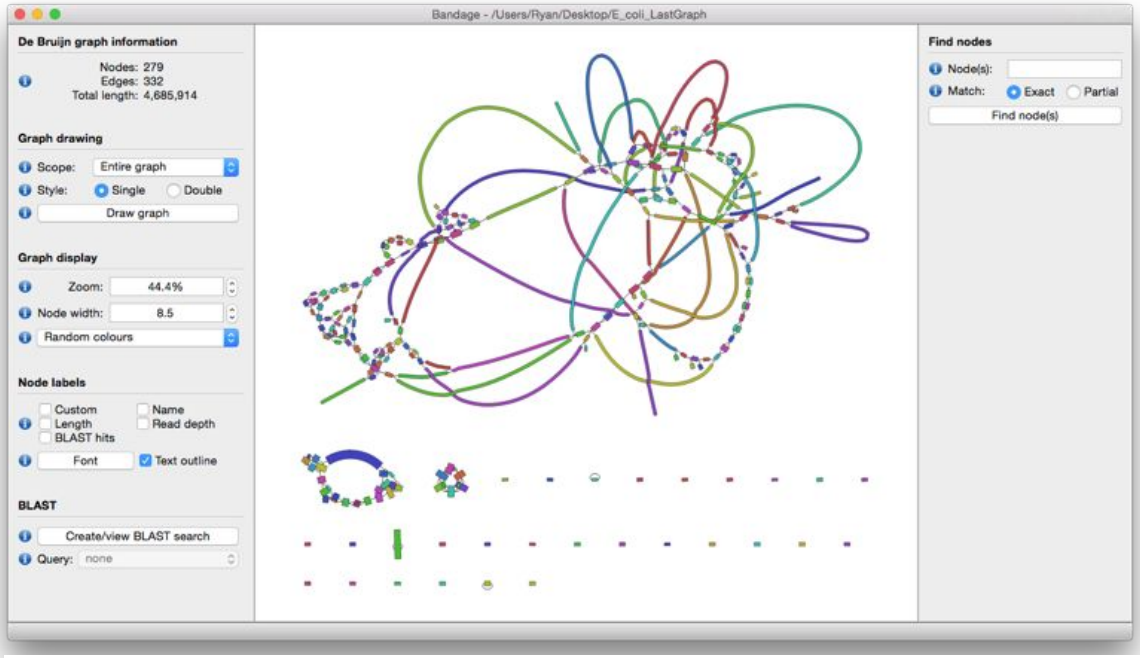
Haplotype phasing via read-pairs, linked-reads, and/or haplotagged BAMs.



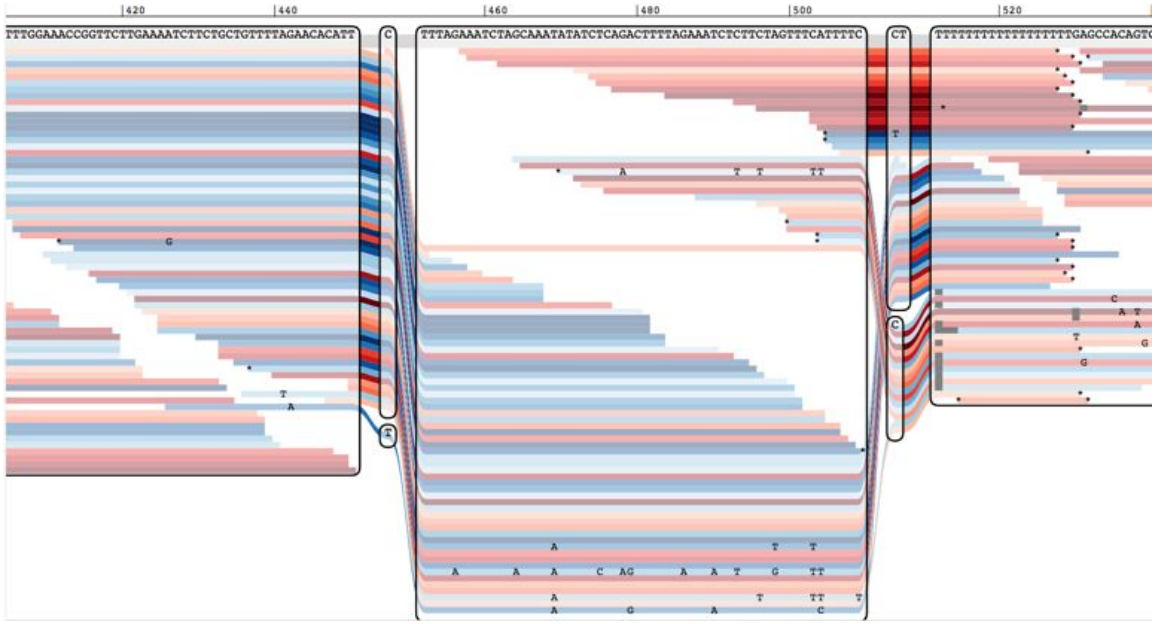
Optimized algorithms for calling of mid-range (50-500bp) mutations and complex events.

Aim 2: *graph serialization and visualization*

Serialization to GFA v1/v2 to support graph visualization in Bandage in a way that users can explore the underlying data with high-resolution and fidelity.



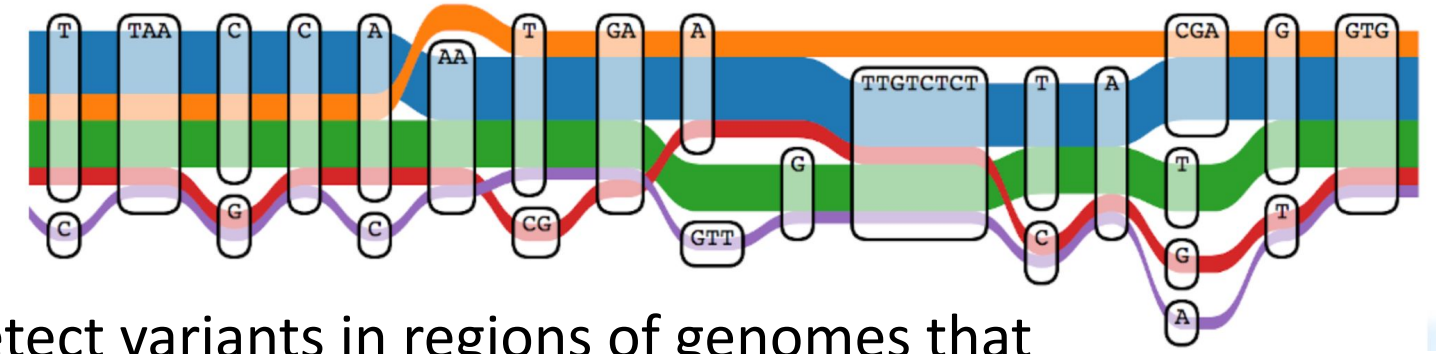
Bandage user interface. Multiple drop-down menus are available to, for example, move, label and color nodes.



SequenceTubeMap graph layout. Sequence graph illustration of a stretch of the BRCA1 gene with known variants from the 1000 Genomes project and supporting reads aligned.

Aim 3: *Cancer genome graphs*

- ***Goal**: Extend Lancet to allow the joint, longitudinal analysis of multiple (>2) cancer genomes*
- **Scenario**: joint analysis of cancer mutations across a large cohort of cancer samples, e.g. The Cancer Genome Atlas (TGCA).



- **Benefits**:
  - *Reduced reference bias*: detect variants in regions of genomes that substantially differ from the reference sequence.
  - *Increase sensitivity* to discover shared events across cancer samples
  - More accurate variant *allele fraction estimates*, critical to understanding sub-clonal structures.
  - *Harmonized representation of variants*: effectively removing the (complicated) step of merging variants independently called in each sample.

Thank you

Do you have interesting datasets to analyze?
We are looking for collaborations!