Practical Considerations of Genomic Data

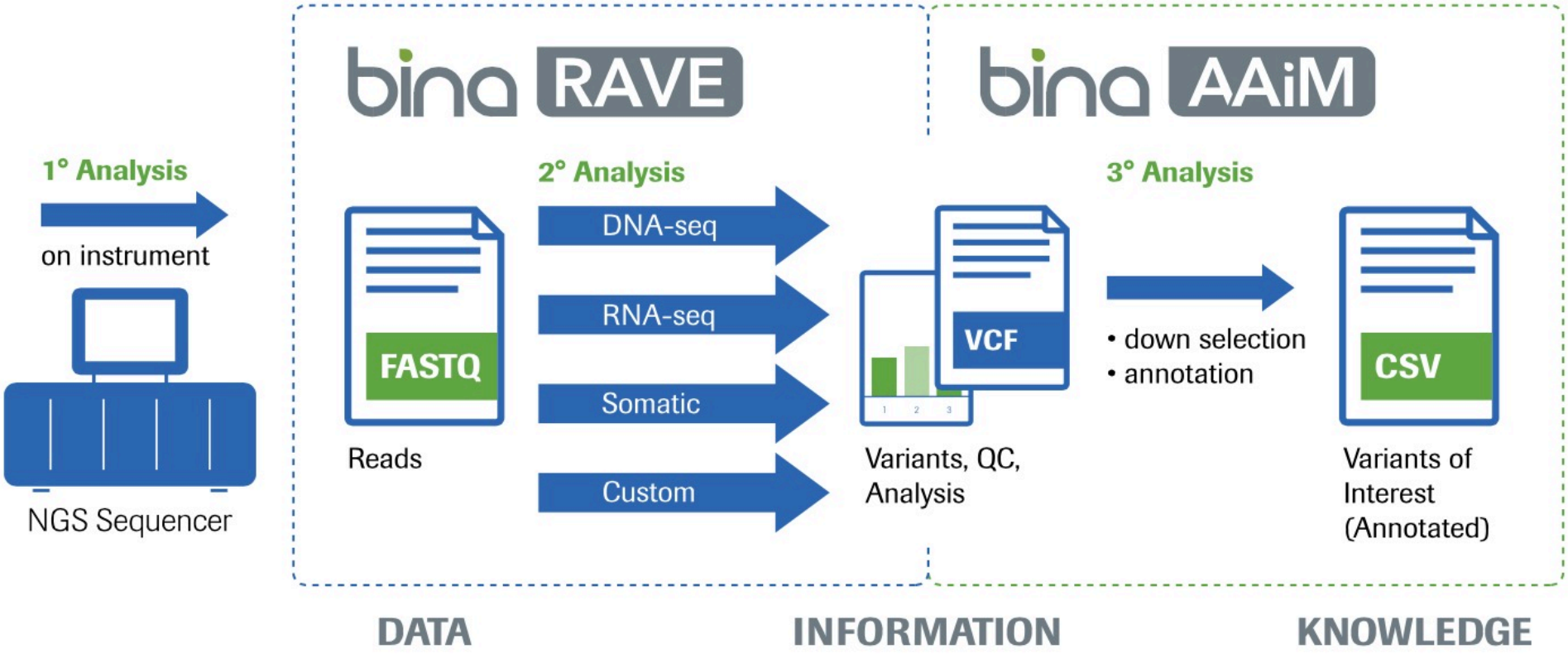
Bayo Lau
Jian Li
Hugo Y. K. Lam

Bina Technologies

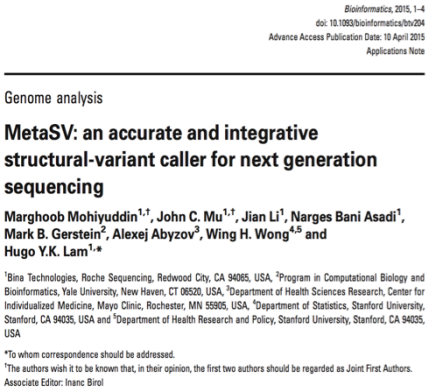
Panel on Genomic Information Compression
MPEG 114

Introduction

What we do



Collaborative scientific innovation



Bioinformatics, 2015, 1–4
doi: 10.1093/bioinformatics/btv204
Advance Access Publication Date: 10 April 2015
Applications Note

Genome analysis

MetaSV: an accurate and integrative structural-variant caller for next generation sequencing

Marghoob Mohiyuddin[1,†], John C. Mu[1,†], Jian Li[1], Narges Bani Asadi[1], Mark B. Gerstein[2], Alexej Abyzov[3], Wing H. Wong[4,5] and Hugo Y.K. Lam[1,*]

[1]Bina Technologies, Roche Sequencing, Redwood City, CA 94065, USA, [2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA, [3]Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA, [4]Department of Statistics, Stanford University, Stanford, CA 94035, USA and [5]Department of Health Research and Policy, Stanford University, Stanford, CA 94035, USA

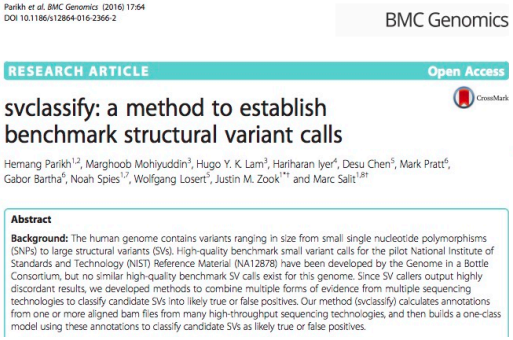
*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Inanc Birol

ARTICLE

Received 3 Oct 2014 | Accepted 21 Apr 2015 | Published 1 Jun 2015 DOI: 10.1038/ncomms8256

Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms

Alexej Abyzov[1,2,3], Shantao Li[2,4], Daniel Rhee Kim[4], Marghoob Mohiyuddin[5], Adrian M. Stütz[6], Nicholas F. Parrish[7], Xinmeng Jasmine Mu[2,3], Wyatt Clark[2,3], Ken Chen[8], Matthew Hurles[9], Jan O. Korbel[6,10], Hugo Y.K. Lam[5], Charles Lee[11] & Mark B. Gerstein[2,3,4]



Parikh *et al. BMC Genomics* (2016) 17:64
DOI 10.1186/s12864-016-2366-2

BMC Genomics

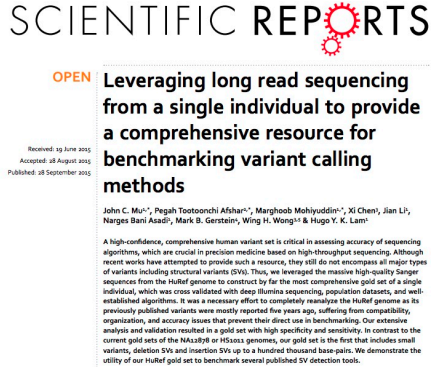
RESEARCH ARTICLE Open Access

svclassify: a method to establish benchmark structural variant calls

Hemang Parikh[1,2], Marghoob Mohiyuddin[3], Hugo Y. K. Lam[3], Hariharan Iyer[4], Desu Chen[5], Mark Pratt[6], Gabor Bartha[6], Noah Spies[1,7], Wolfgang Losert[5], Justin M. Zook[*†] and Marc Salit[1,8†]

Abstract

Background: The human genome contains variants ranging in size from small single nucleotide polymorphisms (SNPs) to large structural variants (SVs). High-quality benchmark small variant calls for the pilot National Institute of Standards and Technology (NIST) Reference Material (NA12878) have been developed by the Genome in a Bottle Consortium, but no similar high-quality benchmark SV calls exist for this genome. Since SV callers output highly discordant results, we developed methods to combine multiple forms of evidence from multiple sequencing technologies to classify candidate SVs into likely true or false positives. Our method (svclassify) calculates annotations from one or more aligned bam files from many high-throughput sequencing technologies, and then builds a one-class model using these annotations to classify candidate SVs as likely true or false positives.



SCIENTIFIC REPORTS

OPEN

Leveraging long read sequencing from a single individual to provide a comprehensive resource for benchmarking variant calling methods

Received: 19 June 2015
Accepted: 28 August 2015
Published: 28 September 2015

John C. Mu[1,*], Pegah Tootoonchi Afshar[1,*], Marghoob Mohiyuddin[1,*], Xi Chen[1], Jian Li[1], Narges Bani Asadi[1], Mark B. Gerstein[1], Wing H. Wong[1,2] & Hugo Y. K. Lam[1]

A high-confidence, comprehensive human variant set is critical in assessing accuracy of sequencing algorithms, which are crucial in precision medicine based on high-throughput sequencing. Although recent works have attempted to provide such a resource, they still do not encompass all major types of variants including structural variants (SVs). Thus, we leveraged the massive high-quality Sanger sequences from the HuRef genome to construct by far the most comprehensive gold set of a single individual, which was cross validated with deep Illumina sequencing, population datasets, and well-established algorithms. It was a necessary effort to completely reanalyze the HuRef genome as its previously published variants were mostly reported five years ago, suffering from compatibility, organization, and accuracy issues that prevent their direct use in benchmarking. Our extensive analysis and validation resulted in a gold set with high specificity and sensitivity. In contrast to the current gold sets of the NA12878 or HS1011 genomes, our gold set is the first that includes small variants, deletion SVs and insertion SVs up to a hundred thousand base-pairs. We demonstrate the utility of our HuRef gold set to benchmark several published SV detection tools.

BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 18 2012, pages 2366–2373
doi:10.1093/bioinformatics/bts450

Sequence analysis

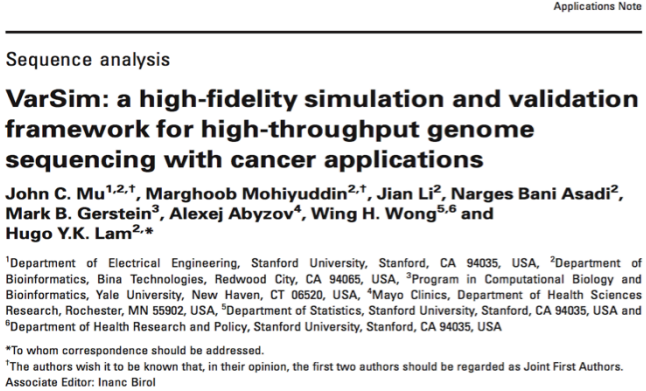
Advance Access publication July 18, 2012

Fast and accurate read alignment for resequencing

John C. Mu[1], Hui Jiang[2], Amirhossein Kiani[3], Marghoob Mohiyuddin[3], Narges Bani Asadi[3] and Wing H. Wong[4,*]

[1]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA, [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA, [3]Bina Technologies Inc., Redwood City and [4]Department of Statistics, Stanford University, Stanford, CA 94305, USA

Associated Editor: Michael Brudno



Bioinformatics, 31(9), 2015, 1469–1471
doi: 10.1093/bioinformatics/btu828
Advance Access Publication Date: 17 December 2014
Applications Note

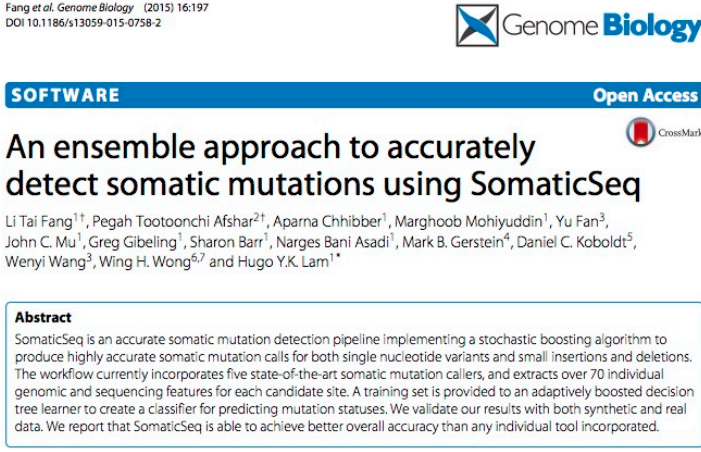
Sequence analysis

VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications

John C. Mu[1,2,†], Marghoob Mohiyuddin[2,†], Jian Li[2], Narges Bani Asadi[2], Mark B. Gerstein[3], Alexej Abyzov[4], Wing H. Wong[5,6] and Hugo Y.K. Lam[2,*]

[1]Department of Electrical Engineering, Stanford University, Stanford, CA 94035, USA, [2]Department of Bioinformatics, Bina Technologies, Redwood City, CA 94065, USA, [3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA, [4]Mayo Clinics, Department of Health Sciences Research, Rochester, MN 55902, USA, [5]Department of Statistics, Stanford University, Stanford, CA 94035, USA and [6]Department of Health Research and Policy, Stanford University, Stanford, CA 94035, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Inanc Birol



Fang *et al. Genome Biology* (2015) 16:197
DOI 10.1186/s13059-015-0758-2

Genome Biology

SOFTWARE Open Access

An ensemble approach to accurately detect somatic mutations using SomaticSeq

Li Tai Fang[1†], Pegah Tootoonchi Afshar[2†], Aparna Chhibber[1], Marghoob Mohiyuddin[1], Yu Fan[3], John C. Mu[1], Greg Gibeling[1], Sharon Barr[1], Narges Bani Asadi[1], Mark B. Gerstein[4], Daniel C. Koboldt[5], Wenyi Wang[3], Wing H. Wong[6,7] and Hugo Y.K. Lam[1*]

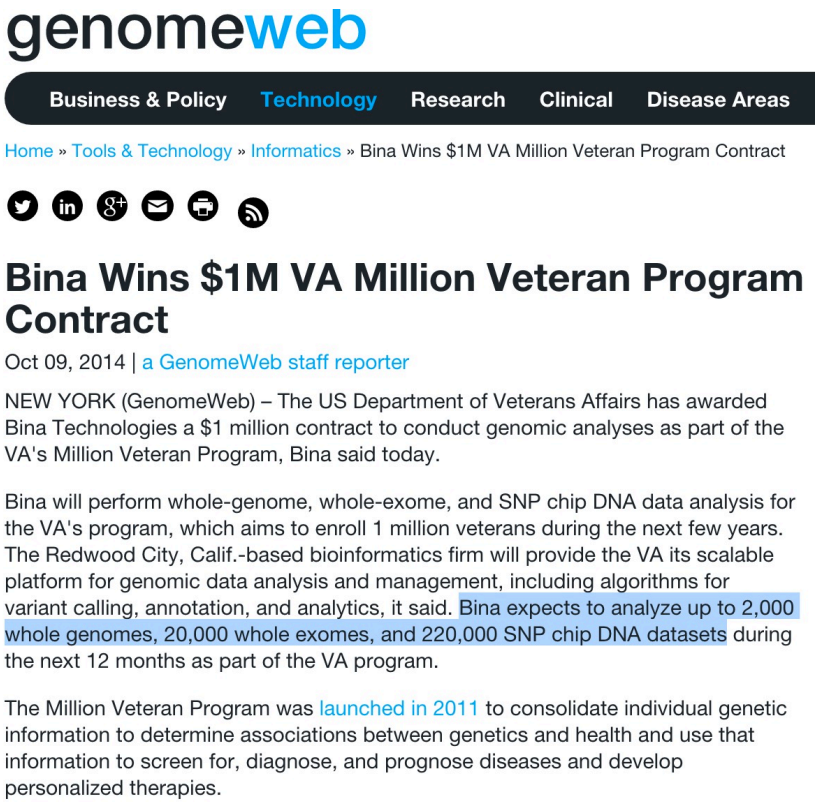
Abstract

SomaticSeq is an accurate somatic mutation detection pipeline implementing a stochastic boosting algorithm to produce highly accurate somatic mutation calls for both single nucleotide variants and small insertions and deletions. The workflow currently incorporates five state-of-the-art somatic mutation callers, and extracts over 70 individual genomic and sequencing features for each candidate site. A training set is provided to an adaptively boosted decision tree learner to create a classifier for predicting mutation statuses. We validate our results with both synthetic and real data. We report that SomaticSeq is able to achieve better overall accuracy than any individual tool incorporated.

Data Volume

Consider: Multi-sample Analysis

- Single-sample
  - limited interpretability
- Multi-sample
  - facilitates study of family or population
  - Rare variants, de novo mutations, etc
- Academic and commercial needs



genomeweb

Business & Policy | Technology | Research | Clinical | Disease Areas

Home » Tools & Technology » Informatics » Bina Wins $1M VA Million Veteran Program Contract

Bina Wins $1M VA Million Veteran Program Contract

Oct 09, 2014 | a GenomeWeb staff reporter

NEW YORK (GenomeWeb) – The US Department of Veterans Affairs has awarded Bina Technologies a $1 million contract to conduct genomic analyses as part of the VA's Million Veteran Program, Bina said today.

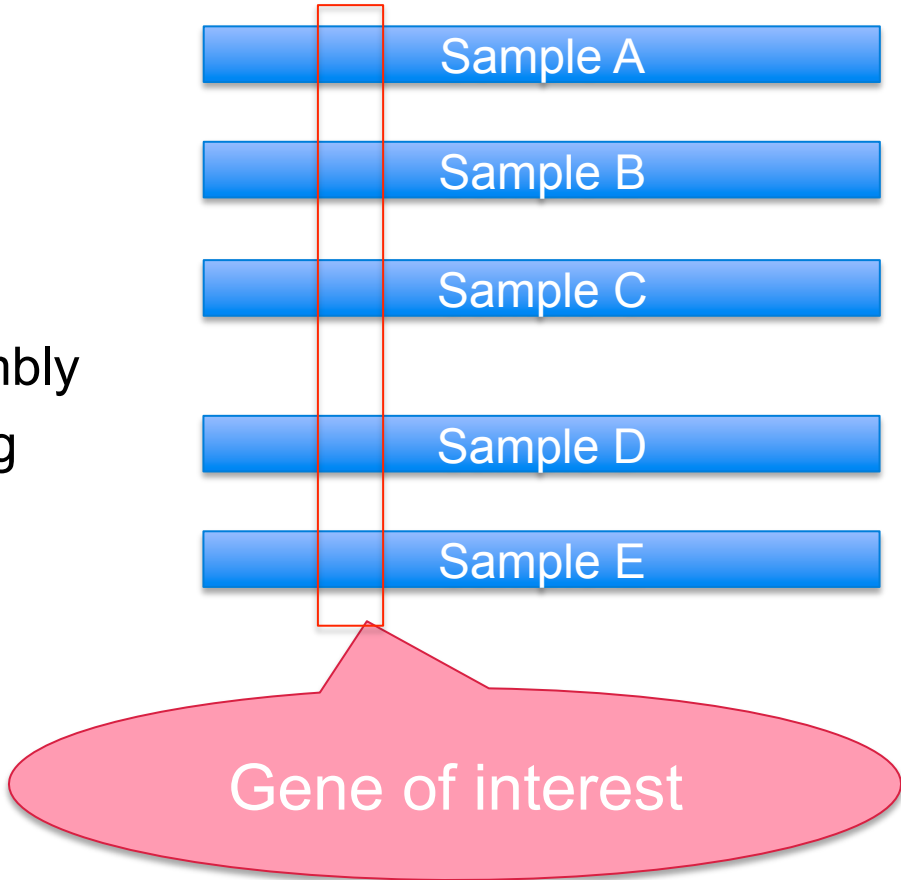
Bina will perform whole-genome, whole-exome, and SNP chip DNA data analysis for the VA's program, which aims to enroll 1 million veterans during the next few years. The Redwood City, Calif.-based bioinformatics firm will provide the VA its scalable platform for genomic data analysis and management, including algorithms for variant calling, annotation, and analytics, it said. Bina expects to analyze up to 2,000 whole genomes, 20,000 whole exomes, and 220,000 SNP chip DNA datasets during the next 12 months as part of the VA program.

The Million Veteran Program was launched in 2011 to consolidate individual genetic information to determine associations between genetics and health and use that information to screen for, diagnose, and prognose diseases and develop personalized therapies.

bina

Consider: Multi-sample Analysis

- Natural Storage
  - Per-sample, sorted by loci
- Cross-sample analysis
  - Trio/population variant-calling/assembly
  - Tumor-normal somatic variant calling
  - Region of interest within a cohort



bina

Consider: Multi-sample Analysis

- Data Retrieval
  - Rapid
  - Small footprint

- Standard to connect data producer and consumer
  - Scalable transfer
  - Meta Data

[Example] https://github.com/ga4gh/schemas

Global Alliance
for Genomics & Health

Schemas for the Data Working Group

bina

#Code2Cure

Data Complexity

bina.com

Consider: TGS

- Single molecule sequencing
- New data characteristics
  - New data format
  - New methodology
  - New discovery

Ross *et al. Genome Biology* 2013, **14**:R51
http://genomebiology.com/2013/14/5/R51

Genome Biology

RESEARCH Open Access

Characterizing and measuring bias in sequence data

Michael G Ross*, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum and David B Jaffe

Abstract

Background: DNA sequencing technologies deviate from the ideal uniform distribution of reads. These biases impair scientific and medical applications. Accordingly, we have developed computational methods for discovering, describing and measuring bias.

Method

Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene

Erick W. Loomis,[1,5] John S. Eid,[2,5] Paul Peluso,[2] Jun Yin,[1] Luke Hickey,[2] David Rank,[2] Sarah McCalmon,[2] Randi J. Hagerman,[3,4] Flora Tassone,[1,4] and Paul J. Hagerman[1,4,6]

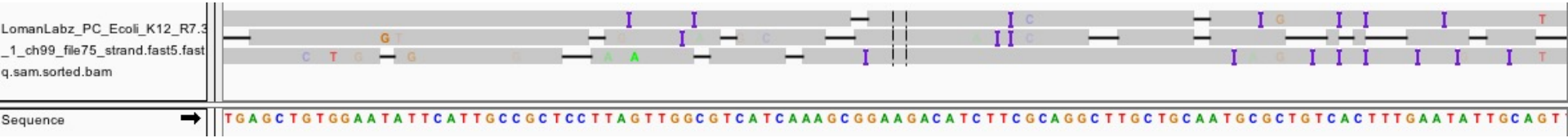
[1]Department of Biochemistry and Molecular Medicine, University of California, Davis, School of Medicine, Davis, California 95616, USA; [2]Pacific Biosciences, Inc., Menlo Park, California 94025, USA; [3]Department of Pediatrics, University of California, Davis, School of Medicine, Sacramento, California 95817, USA; [4]MIND Institute, University of California Davis Medical Center, Sacramento, California 95817, USA

The human fragile X mental retardation I (*FMRI*) gene contains a $(CGG)_n$ trinucleotide repeat in its 5′ untranslated region (5′UTR). Expansions of this repeat result in a number of clinical disorders with distinct molecular pathologies, including fragile X syndrome (FXS; full mutation range, greater than 200 CGG repeats) and fragile X–associated tremor/ataxia syndrome (FXTAS; premutation range, 55–200 repeats). Study of these diseases has been limited by an inability to sequence expanded CGG repeats, particularly in the full mutation range, with existing DNA sequencing technologies. Single-molecule, real-time (SMRT) sequencing provides an approach to sequencing that is fundamentally different from

bina

Consider: TGS

- Much longer reads
- Higher ins/del error rate



- CIGAR string not as effective:

```
2D8M1D4M2D4M1D11M1D10M2D1M1D7M1D11M1I18M1D
```

bina

A complete bacterial genome assembled *de novo* using only nanopore sequencing data

Nicholas J Loman, Joshua Quick & Jared T Simpson

Affiliations | Contributions | Corresponding author

Nature Methods **12**, 733–735 (2015) | doi:10.1038/nmeth.3444
Received 11 March 2015 | Accepted 22 May 2015 | Published online 15 June 2015

Consider: TGS

- Quantify dominant error mode – beyond QV data
- Needed far away from primary analysis

A complete bacterial genome assembled *de novo* using only nanopore sequencing data

Nicholas J Loman, Joshua Quick & Jared T Simpson

Affiliations | Contributions | Corresponding author

Nature Methods **12**, 733–735 (2015) | doi:10.1038/nmeth.3444
Received 11 March 2015 | Accepted 22 May 2015 | Published online 15 June 2015

Assembly and diploid architecture of an individual human genome via single-molecule technologies

Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, Heng Dai, Markus Hsi-Yang Fritz, Han Cao, Ariella Cohain, Gintaras Deikus, Russell E Durrett, Scott C Blanchard, Roger Altman, Chen-Shan Chin, Yan Guo, Ellen E Paxinos, Jan O Korbel, Robert B Darnell, W Richard McCombie, Pui-Yan Kwok, Christopher E Mason, Eric E Schadt & Ali Bashir – Show fewer authors

Affiliations | Contributions | Corresponding author

Nature Methods **12**, 780–786 (2015) | doi:10.1038/nmeth.3454
Received 05 December 2014 | Accepted 28 May 2015 | Published online 29 June 2015

Assembling large genomes with single-molecule sequencing and locality-sensitive hashing

Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin & Adam M Phillippy

Affiliations | Contributions | Corresponding author

Nature Biotechnology **33**, 623–630 (2015) | doi:10.1038/nbt.3238
Received 14 August 2014 | Accepted 08 April 2015 | Published online 25 May 2015 | Corrected online 06 October 2015
Corrigendum (October, 2015)

Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data

Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner & Jonas Korlach

Affiliations | Contributions | Corresponding author

Nature Methods **10**, 563–569 (2013) | doi:10.1038/nmeth.2474
Received 31 January 2013 | Accepted 04 April 2013 | Published online 05 May 2013

bina

For Res

Consider: TGS

- Oxford nanopore
  - FASTQ
  - 5-mer
  - probabilities (floating point)
  - Signal strength (floating point)
  - Signal duration (floating point)

Jared Simpson, Ontario Institute for Cancer Research, private communication

user mw55309, https://github.com/mw55309/PoreCamp/blob/master/PoreCamp.md

Poretools: a toolkit for analyzing nanopore sequence data

Nicholas J. Loman[1,*] and Aaron R. Quinlan[2,*]

Author Affiliations

[1]Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK and [2]Department of Public Health Sciences, University of Virginia, Charlottesville 22932, VA, USA

*To whom correspondence should be addressed.

Received July 29, 2014.
Revision received July 29, 2014.
Accepted August 11, 2014.

bina

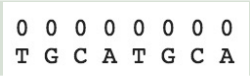
Consider: TGS

- Some per-base data by Pacific Biosciences
  - base call
  - QV
  - QV for insertion
  - QV for deletion
  - QV for substitution
  - QV for merge
  - Most probable deleted base
  - Most probable substitution base
  - Inter-pulse duration IPD (16-bit unsigned)
  - Pulse width PW (16-bit unsigned)

[v2] http://files.pacb.com/software/instrument/2.0.0/bas.h5%20Reference%20Guide.pdf
[v2] http://pacbiofileformats.readthedocs.org/en/3.0/CmpH5Spec.html
[v3] https://github.com/PacificBiosciences/PacBioFileFormats/blob/3.0/BAM.rst

bina

Consider: TGS

Version	PacBio v2	PacBio v3
Unmapped reads	h5	BAM (per-base data in optional fields)
Mapped reads	h5	BAM
Alignment	8-bit-per-base 	SEQ, REF, POS, CIGAR
Meta data	Exhaustive, well organized	Encoded in header/tags

[v2] http://files.pacb.com/software/instrument/2.0.0/bas.h5%20Reference%20Guide.pdf
[v2] http://pacbiofileformats.readthedocs.org/en/3.0/CmpH5Spec.html
[v3] https://github.com/PacificBiosciences/PacBioFileFormats/blob/3.0/BAM.rst

bina

#Code2Cure

Data Usability

bina.com

Consider: Usability (merely my 2 cents)

- Adoption
  - “Big shots” have huge influence
  - Licensing is a critical bridge between academic and industry

- Application program interface
  - Cross-language, cross-platform, Unsigned integer not easily universal
  - Distribution
  - HTSLIB (C) and HTSJDK (Java) are already built into many production infrastructures

- Extendibility and backward compatibility
  - Component-by-component adoption of a stable pipeline

bina

Consider: Usability (merely my 2 cents)

- Access pattern
  - Partial retrieval
  - Multi-key sorting and grouping

- stdin/stdout
  - piping
  - I/O and network overhead

- Format validator
  - easy enforcement
  - hints for correction

bina

Considerations

- Data Volume
- Data complexity
- Usability

Acknowledgements

- Jian Li, Senior Bioinformatics Scientist
- LiTai Fang, Senior Bioinformatics Scientist
- Lijing Yao, Bioinformatics Scientist

- Marghoob Mohiyuddin, Staff Scientist
- Aparna Chhibber, Senior Scientist
- John C. Mu, Research Engineer

- Hugo Y. K. Lam, Senior Director of Bioinformatics

bina