Long insert sequencing of 1000 Genomes Yoruban trio at JAX-GM

Ankit Malhotra, Xiaoan Ruan, Chengsheng Zhang,
Jonas Korlach, Charles Lee

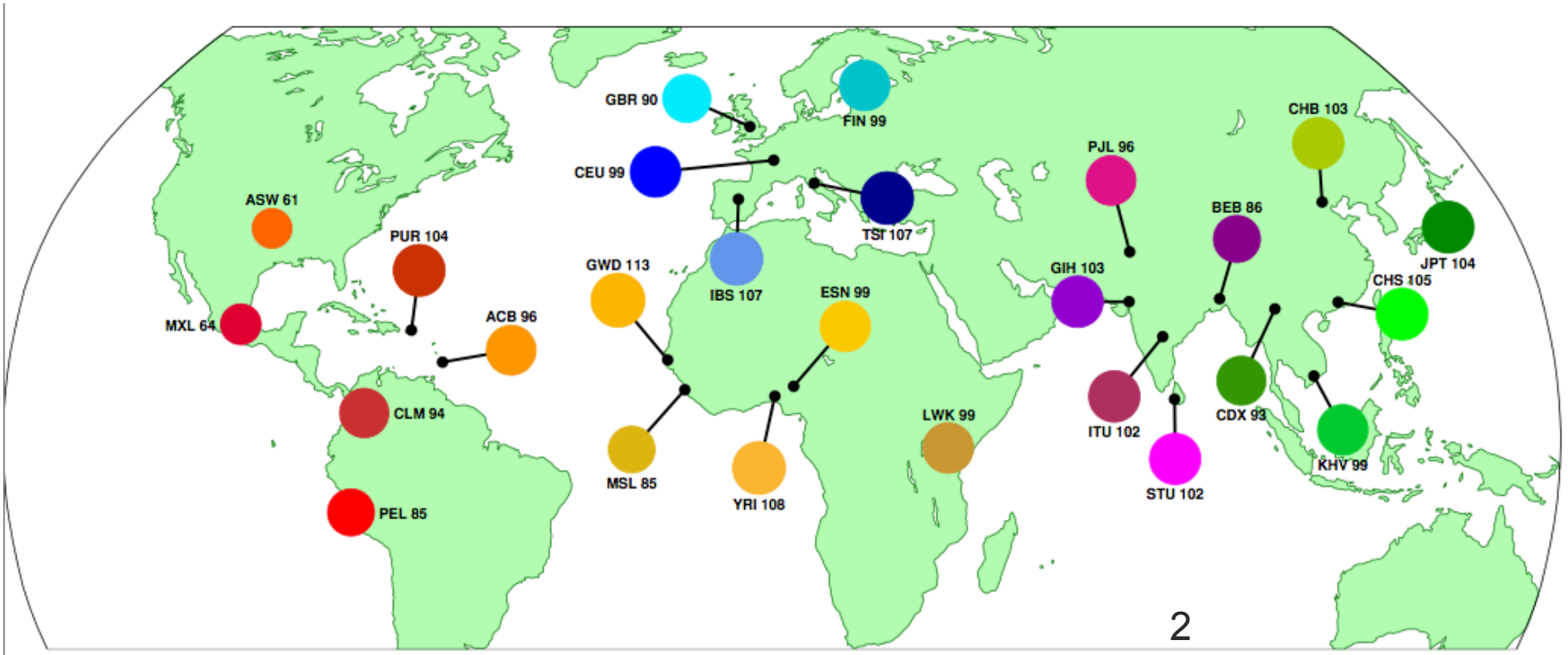
The Jackson Laboratory
Leading the search for tomorrow's cures

1000 Genomes

A Deep Catalog of Human Genetic Variation

1000 Genomes is an international collaboration of researchers around the world. The central aim of the 1000 Genomes Project structural variation subgroup is to comprehensively identify the landscape of genetic structural variations (SV) affecting human health and diversity

Over the various phases the project has grown to include data from 2500+ individuals from across 26 different world populations.



**slide provided by Eugene Gardner*

1000 Genomes Project, Phase 4

As part of next and final phase of the 1000G Structural variation project, we plan to delve deeper into a small number of selected genomes by multiple cutting edge and high-resolution technologies and methodologies.

Selected trios :

- Yoruban Trio
  - NA19238
  - NA19239
  - NA19240
- Han Chinese Trio
  - HG00512
  - HG00513
  - HG00514
- Puerto Rican Trio
  - HG00731
  - HG00732
  - HG00733

Platforms / Technologies :

- PacBio SMRT long insert sequencing
- Illumina PCR-Free short read sequencing
- Illumina Moleculo long read sequencing
- Bio Nano Genomics optical mapping
- 5kb Jumping libraries
- Mate-pair sequencing
- Fosmid sequencing
- 3D Structure mapping using HiC libraries
- Affy CytoScan HD array platform

We expect that the proposed research activities will lead to development of new experimental and analysis methodologies for assessing the landscape of structural variations genome wide. In addition, the set of SV calls generated by this project will provide a valuable resource of high confidence and validated germline structural variation loci to the general research community

1000 Genomes Project, Phase 4

As part of next and final phase of the 1000G Structural variation project, we plan to delve deeper into a small number of selected genomes by multiple cutting edge and high-resolution technologies and methodologies.

Selected trios :

- YRI Trio
  - NA19238
  - NA19239
  - NA19240
- Han Chinese Trio
  - HG00512
  - HG00513
  - HG00514
- Puerto Rican Trio
  - HG00731
  - HG00732
  - HG00733

Platforms / Technologies :

- PacBio SMRT long insert sequencing
- Illumina PCR-Free short read sequencing
- Illumina Moleculo long read sequencing
- Bio Nano Genomics optical mapping
- 5kb Jumping libraries
- Mate-pair sequencing
- Fosmid sequencing
- 3D Structure mapping using HiC libraries
- Affy CytoScan HD array platform

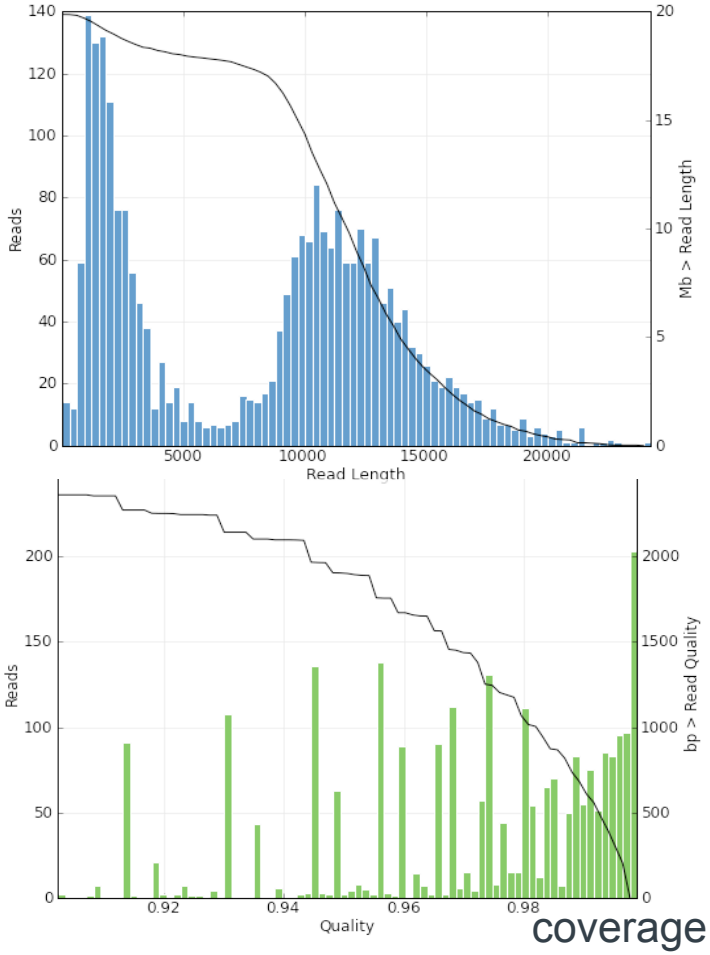
We expect that the proposed research activities will lead to development of new experimental and analysis methodologies for assessing the landscape of structural variations genome wide. In addition, the set of SV calls generated by this project will provide a valuable resource of high confidence and validated germline structural variation loci to the general research community

Pacific Biosciences SMRT sequencing platform

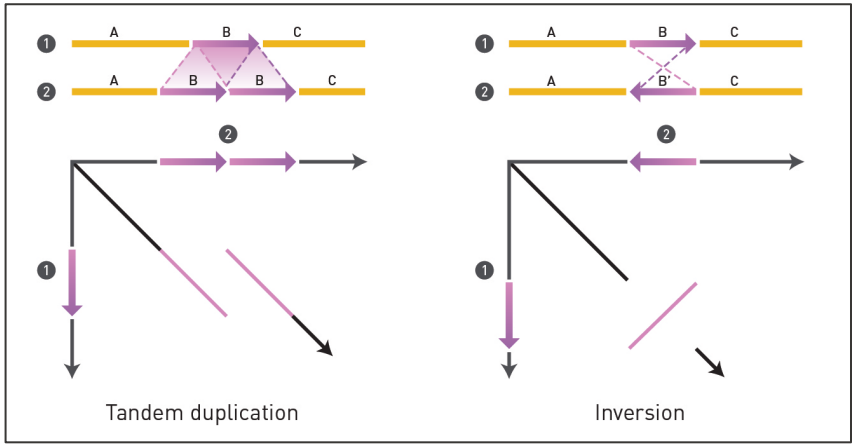
Pacific Biosciences SMRT sequencing platform allows for sequencing long and complex regions of the genome with an average read size in the excess of 10kb.

Pacific Biosciences SMRT sequencing platform

Pacific Biosciences SMRT sequencing platform allows for sequencing long and complex regions of the genome with an average read size in the excess of 10kb.



Examples of SV detected by long read data

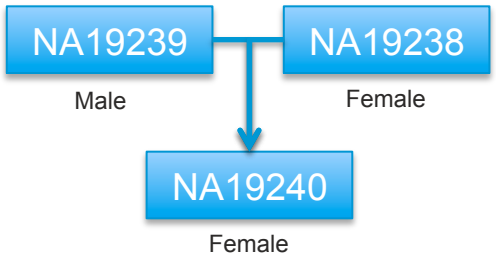


www.pacificbiosciences.com

Example of read length / quality for one run on the Pacbio machine

For each of the identified trios, we aim to generate 40x coverage data for the child and 20x coverage data for the parents

Deep Coverage PacBio sequencing of a 1000 Genomes Yoruban Trio



Sample	In-house Sequenced Datasets
NA19238	115
NA19239	108
NA19240	207

1000G Trio_Child sample (31.7x)	207 datasets					
	reads	mean DNA pol readlength bp	QV	total Mb	mean insert readlength bp	Mean insert QV
Total reads# & Mb:	9,036,164			95,187		
Per SMRTcell:	43,653	10,606 bp	0.83	460	7,943 bp	0.84
			coverage X:	31.7		

1000G Trio_Mother sample (21.7x)	115 datasets					
	reads	mean DNA pol readlength bp	QV	total Mb	mean insert readlength bp	Mean insert QV
Total reads# & Mb:	6,390,590			64,956		
Per SMRTcell:	55,570	10,026 bp	0.83	565	7,721 bp	0.84
			coverage X:	21.7		

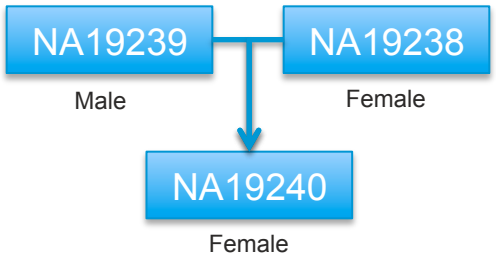
1000G Trio_Father sample (20.5x)	108 datasets					
	reads	mean DNA pol readlength bp	QV	total Mb	mean insert readlength bp	insert QV
Total reads# & Mb:	5,725,433			61,529		
Per SMRTcell:	53,509	10,616 bp	0.83	575	7,976 bp	0.84
			coverage X:	20.5		

*Additional 10x child data sequenced by PacBio

Primary data deposition at EMBL

- Thanks to Laura Clarke and the data submission team at the ENA, all primary data has been uploaded to ENA and made available at the 1000 genomes FTP site
- One tar.gz file per run (NA19240 – 242 runs, NA19239 – 102 runs, NA19238 – 112 runs)
- Each run consists of
  - 1 bas.h5,
  - 3 bax.h5 files and
  - 1 metadata.xml file

Data Pre-processing analysis and data availability



Data pre-processing steps :

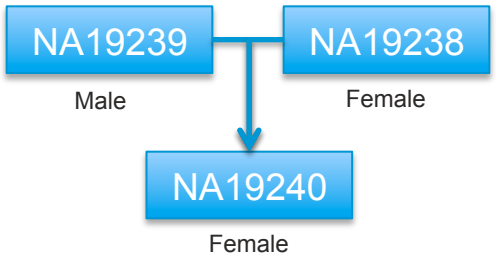
1. Extract sub-reads and reads_of_insert (fastq) for each sequencing run using PacBio SMRT-Portal
2. Alignments of both sub-reads and reads_of_insert to 1000 genomes version of the hg19 genome
    a) Using recommended BLASR parameters

        -unaligned unaligned.fastq -bestn 2 -maxAnchorsPerPosition 100 -advanceExactMatches 10 -affineAlign -affineOpen 100 -affineExtend 0 -insertion 5 -deletion 5 -extend -maxExtendDropoff 20 -clipping subread -header $RFILE -sa $SAFILE -m 5 -clipping soft -out alignmentsRecommended.m5 -nproc 8

    b) Output in both m5 and bam formats (~ 1TB total size of each)
3. Concatenate all alignments from each of the samples (NA19238 / NA19239 / NA19240) into a single combined m5 / bam file

* Whole genome alignment files (m5 and bam) will soon be available at the 1000 Genomes FTP site.

SV Discovery process for a 1000 Genomes Yoruban Trio

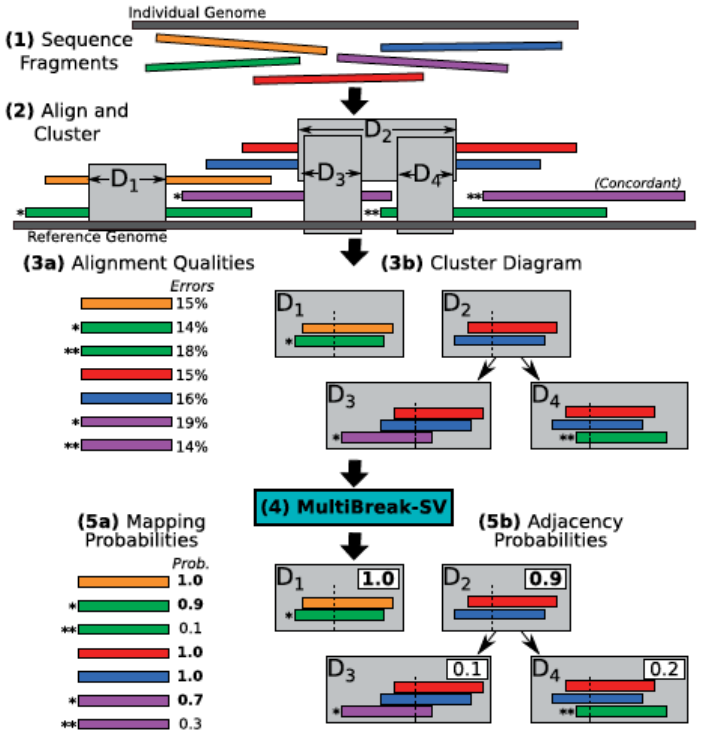


Currently working on an SV discovery workflow:

1. Extract reads and reads_of_insert for each sequencing run
2. Aligned the sub-reads using BLASR (recommended parameters)
3. Merging different runs into a combined file
4. Process combined m5 files with GASV-Pro / MultiBreak-SV to call Structural Variant events
5. **Evaluate results and Improve the methodology to allow for SV discovery from diploid genomes**

Processing using MultiBreak-SV

- **MultiBreak-SV***, is a structural variant caller that allows the user to combine multiple types of sequencing data from different platforms
- Data from different platforms are individually aligned and discordant pairs clustered together to identify putative breakpoints.
- MultiBreak-SV calculates mapping / adjacency probabilities for each breakpoint by aggregating information about alignment qualities and read clusters.

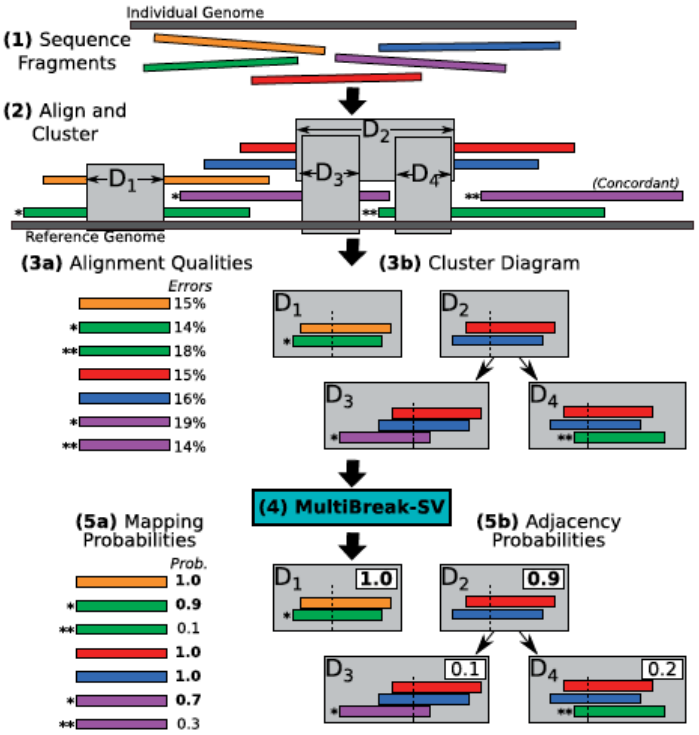


*"*Characterization of Structural Variants with Single Molecule and Hybrid Sequencing Approaches",* Ritz et al, Bioinformatics, Oct 2014

Processing using MultiBreak-SV

- For each individual, both sub-reads and circular consensus reads (CCS) were processed separately with MultiBreak-SV

Type	Sample	All Clusters	#Clusters, size > 1	D	I	T	V
CCS	NA19240	6,046	51	51			
	NA19238	4,259	41	41			
	NA19239	5,509	36	36			
All	NA19240	135,356	1,500	1,498		2	
	NA19238	77,781	497	495	2	0	
	NA19239	66,579	434	433			1



Currently evaluating output from MBSV to evaluate probabilities of each SV event

*"*Characterization of Structural Variants with Single Molecule and Hybrid Sequencing Approaches",* Ritz et al, Bioinformatics, Oct 2014

Future Work

- Improve SV calling by expanding method to handle diploid nature of data
- SNP/Indel calling using GATK
- Combined processing with Illumina paired end sequencing data
- Assembly using Celera / Quiver workflows from PacBio
- Integrate sequence based calls with other SV callsets, such as optical mapping / HiC based 3D maps.

Acknowledgements

Jackson Labs
- Dr. Charles Lee
- Chengsheng Zhang
- Xiaoan Ruan

Other members of the Lee lab and Computational Sciences

PacBio
- Jonas Korlach and team

EMBL/ENA
- Laura Clarke
- Data Submission Team

Brown University
- Benjamin Raphael
- Anna Ritz