



SMRT Link and Analysis Tools for PacBio Data

SMRT Informatics Developers Conference - January 17, 2018

AGENDA

–5.1.0 release goals

–New features

- Support for key analysis applications
 - Multiplexed microbial assembly
 - Structural variant calling
 - De novo assembly
- Barcoding workflow redesign
- Sample Setup redesign
- Data Management and SMRT Analysis usability improvements
- Release of Iso-Seq 2 [Beta]
- Minor Variants support for custom target configs



Release Goals

HIGH-LEVEL GOALS

- Increased system throughput
 - Support for longer movies
 - Chemistry improvements
- End to end support for high-priority applications
- Focus on usability aspects of SMRT Link
- Preliminary R&D work on
 - Building out No-Amp targeted pulldown (Cas9) solution
 - Diploid consensus

SMRT TOOLS - NEW FEATURES

Applications

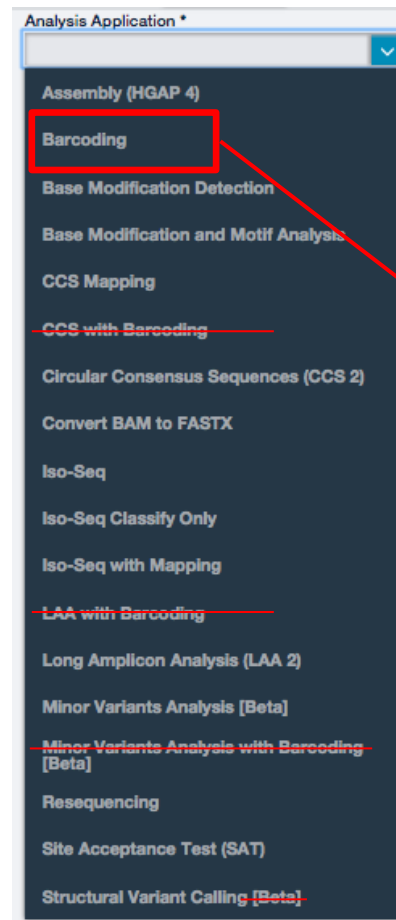
- Multiplexed Microbial Assembly
 - Completely redesigned barcode workflow
 - New barcode calling software
- Structural Variation
 - Support for multi-sample (joint) calling
 - Moved out of Beta status
- De Novo Assembly
 - Support for running unzip after HGAP.4
 - Support for generating GFA output
 - Binary release now available
- Iso-Seq
 - Address scalability issues
- Minor Variants
 - Support for adding custom target configurations (gene annotations)

SMRT LINK – NEW FEATURES

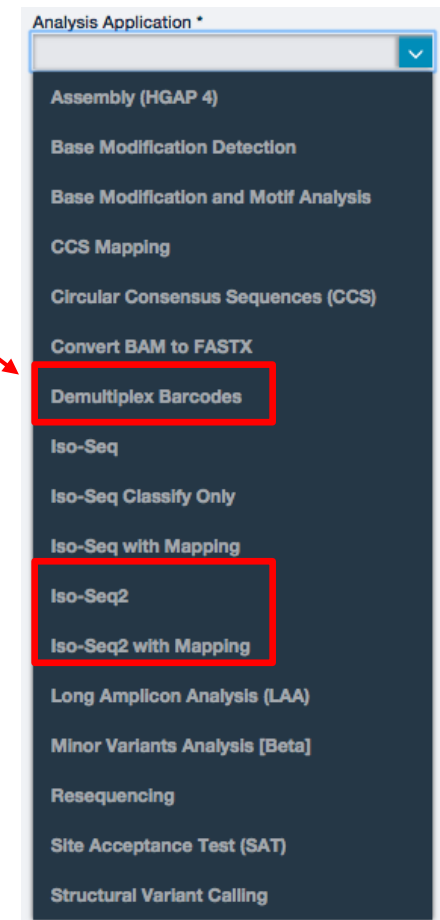
SMRT Link

- Barcoding workflow redesign
- Redesign of Sample Setup
- Usability improvements
 - Data Management
 - SMRT Analysis
- Analysis import/export

5.0.0



5.1.0



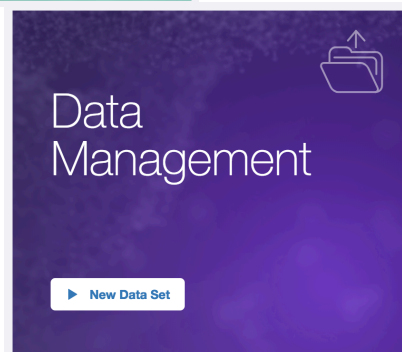


Multiplexed Microbial Assembly

REDESIGN OF BARCODING WORKFLOW



- Specify sample names of individual barcodes

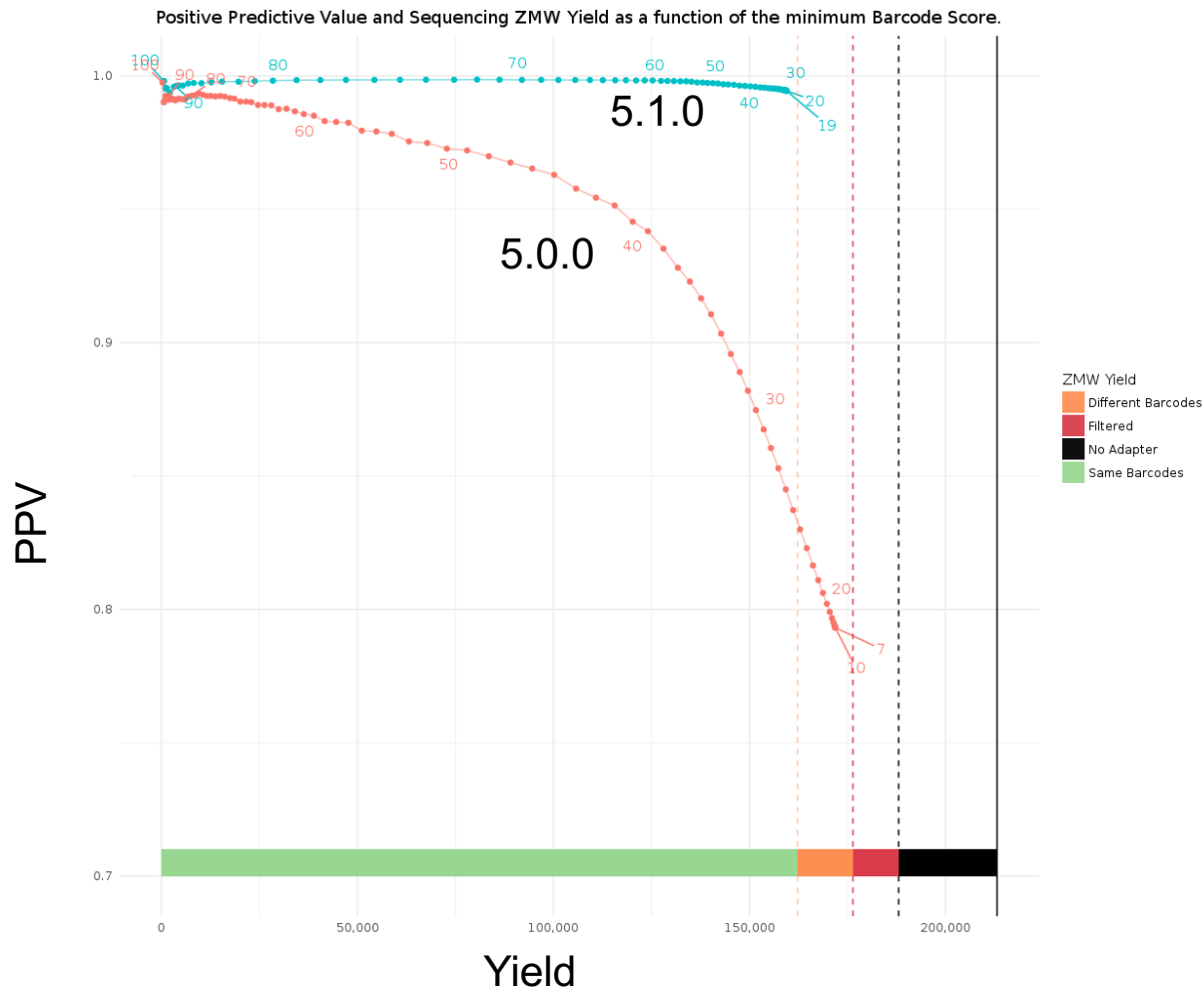


- Better assignment of reads to barcodes
- New, more efficient demultiplexing algorithm
- Extensive new QC metrics

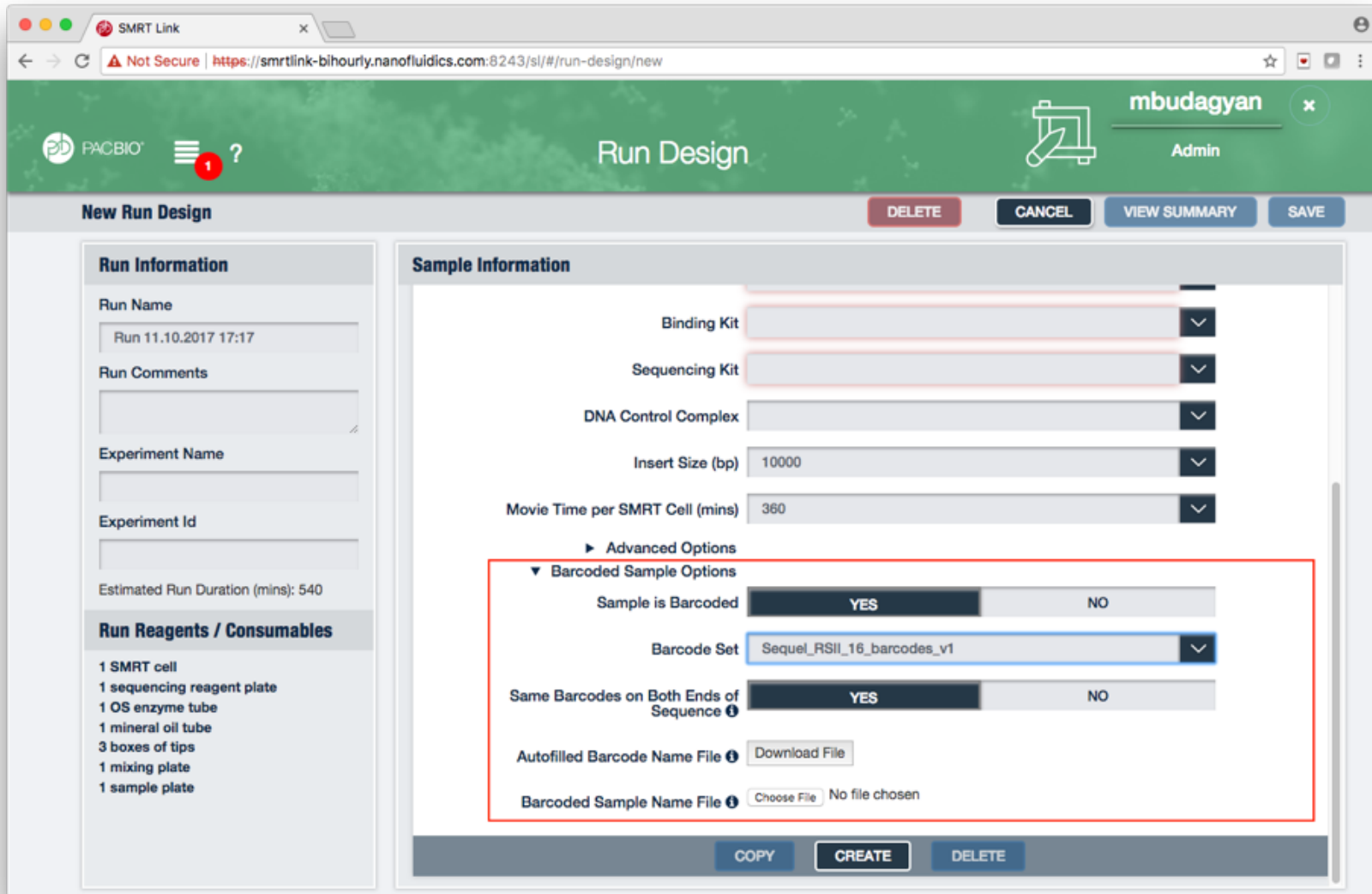


- Flexibility to launch analyses on a single barcode or as a batch with custom parameters

DE-MULTIPLEXING PERFORMANCE – 5.0.0 VS 5.1.0



RUN DESIGN: NEW SECTION FOR BARCODES



The screenshot shows the 'Run Design' interface in a browser window. The page title is 'Run Design' and the user is logged in as 'mbudagyan'. The interface is divided into two main sections: 'Run Information' and 'Sample Information'.

Run Information:

- Run Name: Run 11.10.2017 17:17
- Run Comments: (empty text area)
- Experiment Name: (empty text area)
- Experiment Id: (empty text area)
- Estimated Run Duration (mins): 540

Run Reagents / Consumables:

- 1 SMRT cell
- 1 sequencing reagent plate
- 1 OS enzyme tube
- 1 mineral oil tube
- 3 boxes of tips
- 1 mixing plate
- 1 sample plate

Sample Information:

- Binding Kit: (dropdown menu)
- Sequencing Kit: (dropdown menu)
- DNA Control Complex: (dropdown menu)
- Insert Size (bp): 10000
- Movie Time per SMRT Cell (mins): 360

Advanced Options:

- Barcoded Sample Options (highlighted in red):**
 - Sample is Barcoded: YES (selected)
 - Barcode Set: Sequel_RSII_16_barcode_v1
 - Same Barcodes on Both Ends of Sequence: YES (selected)
 - Autofilled Barcode Name File: Download File
 - Barcoded Sample Name File: Choose File (No file chosen)

Buttons at the bottom of the 'Sample Information' section include COPY, CREATE, and DELETE.

BARCODED SAMPLE NAME FILE

```
Barcode_Names (6).csv
1 Barcode Name,Bio Sample Name (allowed characters:
2 bc1054—bc1054,Alice
3 bc1093—bc1093,Bob
4 bc1004—bc1004,Charles
5 bc1080—bc1080,
6 bc1100—bc1100,
7 bc1109—bc1109,
8 bc1032—bc1032,
9 bc1063—bc1063,
10 bc1002—bc1002,
11 bc1070—bc1070,
12 bc1115—bc1115,
13 bc1016—bc1016,
14 bc1101—bc1101,
15 bc1055—bc1055,
16 bc1118—bc1118,
17 bc1048—bc1048,
18
```

Run Information

Run Name: Run 11.10.2017 17:17

Run Comments:

Experiment Name:

Experiment Id:

Estimated Run Duration (mins): 540

Run Reagents / Consumables

- 1 SMRT cell
- 1 sequencing reagent plate
- 1 OS enzyme tube
- 1 mineral oil tube
- 3 boxes of tips
- 1 mixing plate
- 1 sample plate

Sample Information

Binding Kit:

Sequencing Kit:

DNA Control Complex:

Insert Size (bp): 10000

Movie Time per SMRT Cell (mins): 360

Advanced Options

Barcoded Sample Options

Sample is Barcoded: YES

Barcode Set: Sequel_RSII_16_bar_codes_v1

Same Barcodes on Both Ends of Sequence: YES

Autofilled Barcode Name File: Download File

Barcoded Sample Name File: Choose File No file chosen

COPY CREATE DELETE

- Fill in the sample names for the barcodes used
- Upload the file

QC METRICS IN RESULTS OF DEMULTIPLEX BARCODES

Analysis Results - janet demux test SUCCESSFUL COPY DELETE

Barcodes

	Value	Analysis Metric
Unique Barcodes	95	Unique Barcodes
Barcoded Reads	472,948	Barcoded Reads
Unbarcoded Reads	181,179	Unbarcoded Reads
Mean Reads	4,972	Mean Reads
Max. Reads	7,655	Max. Reads
Min. Reads	2,036	Min. Reads
Mean Read Length	80,527	Mean Read Length
Mean Longest Subread Length	95,376	Mean Longest Subread Length

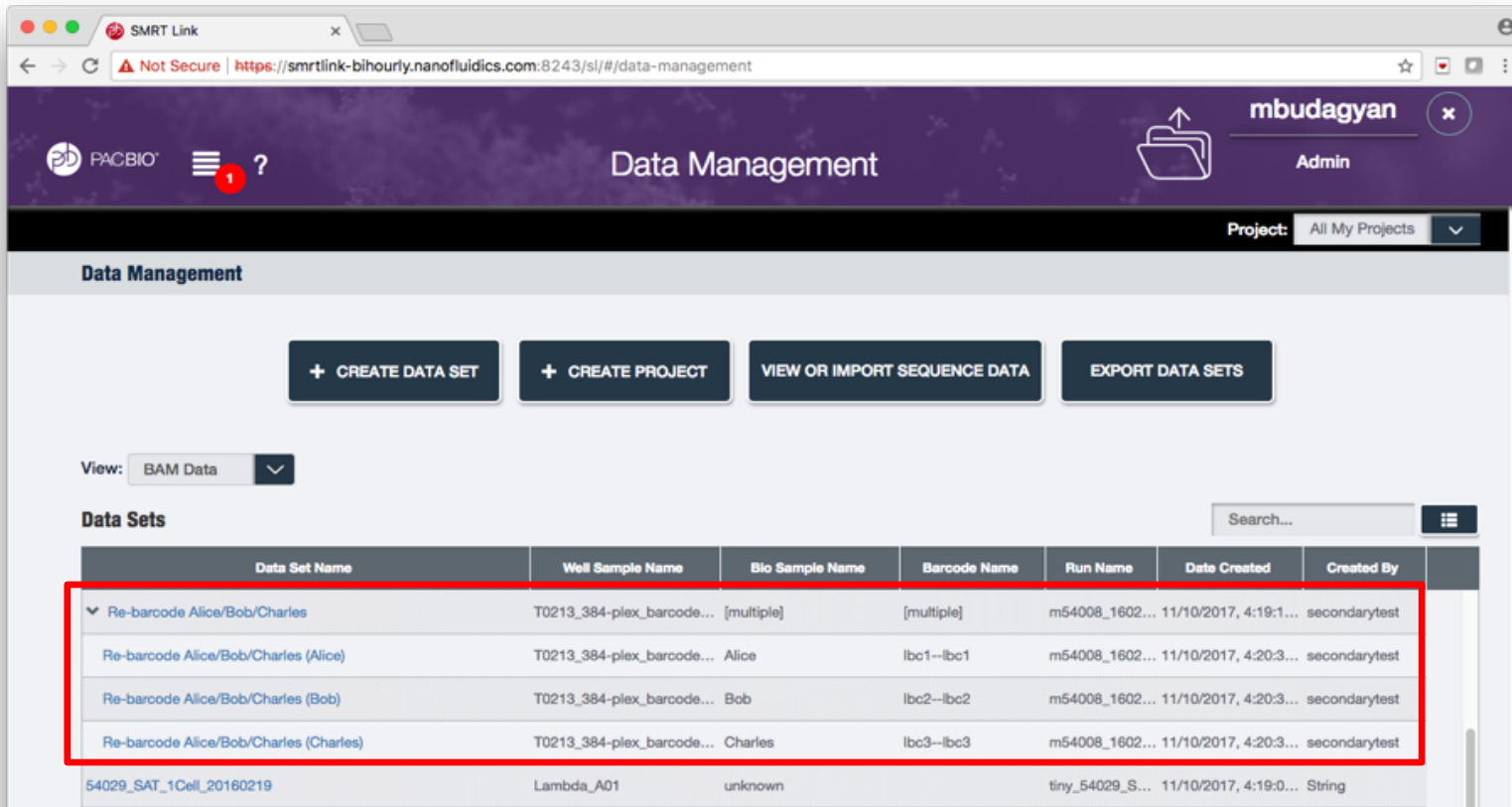
Summary Metrics

- Barcode Data
- Barcoded Read Statistics
- Barcode Quality Scores
- Barcoded Read Binned Histograms

Diagnostic Plots:

- Number Of Reads Per Barcode:** Line graph showing cumulative reads vs. barcode rank.
- Barcode Frequency Distribution:** Histogram of barcode counts.
- Mean Read Length Distribution:** Histogram of mean read lengths.
- Barcode Quality Score Distribution:** Histogram of barcode quality scores.
- Read Length Distribution By Barcode:** Heatmap of read lengths across barcodes.
- Barcode Quality Distribution By Barcode:** Heatmap of quality scores across barcodes.

ONE DATASET GENERATED PER BARCODE IN HIERARCHICAL DISPLAY



The screenshot shows the SMRT Link Data Management interface. The top navigation bar includes the PACBIO logo, a menu icon with a red notification bubble containing the number '1', the title 'Data Management', a folder icon, and the user name 'mbudagyan Admin'. A 'Project' dropdown menu is set to 'All My Projects'. Below the navigation bar, there are four main action buttons: '+ CREATE DATA SET', '+ CREATE PROJECT', 'VIEW OR IMPORT SEQUENCE DATA', and 'EXPORT DATA SETS'. The 'View' dropdown is set to 'BAM Data'. The 'Data Sets' section features a search bar and a table with the following columns: Data Set Name, Well Sample Name, Bio Sample Name, Barcode Name, Run Name, Date Created, and Created By. A red box highlights a hierarchical tree structure for the data set 'Re-barcode Alice/Bob/Charles', which is expanded to show four sub-entries, each with its own row in the table.

Data Set Name	Well Sample Name	Bio Sample Name	Barcode Name	Run Name	Date Created	Created By
Re-barcode Alice/Bob/Charles	T0213_384-plex_barcode...	[multiple]	[multiple]	m54008_1602...	11/10/2017, 4:19:1...	secondarytest
Re-barcode Alice/Bob/Charles (Alice)	T0213_384-plex_barcode...	Alice	lbc1--lbc1	m54008_1602...	11/10/2017, 4:20:3...	secondarytest
Re-barcode Alice/Bob/Charles (Bob)	T0213_384-plex_barcode...	Bob	lbc2--lbc2	m54008_1602...	11/10/2017, 4:20:3...	secondarytest
Re-barcode Alice/Bob/Charles (Charles)	T0213_384-plex_barcode...	Charles	lbc3--lbc3	m54008_1602...	11/10/2017, 4:20:3...	secondarytest
54029_SAT_1Cell_20160219	Lambda_AD1	unknown		tiny_54029_S...	11/10/2017, 4:19:0...	String

- Bio Sample Name assigned by Barcode Name if not provided
 - CSV is optional
- Ability to modify Bio Sample Name and demux parameters

ANALYZE EACH BARCODED DATASET WITH CUSTOM PARAMETERS

Create New Analysis - Settings CANCEL START

Project: All My Projects

Analysis Application *
Assembly (HGAP 4)

Analysis of Multiple Datasets

Analysis Type *
One Analysis on All Data Sets
One Analysis on All Data Sets
 One Analysis per Data Set - Identical Parameters
One Analysis per Data Set - Custom Parameters
 ADVANCED ANALYSIS PARAMETERS

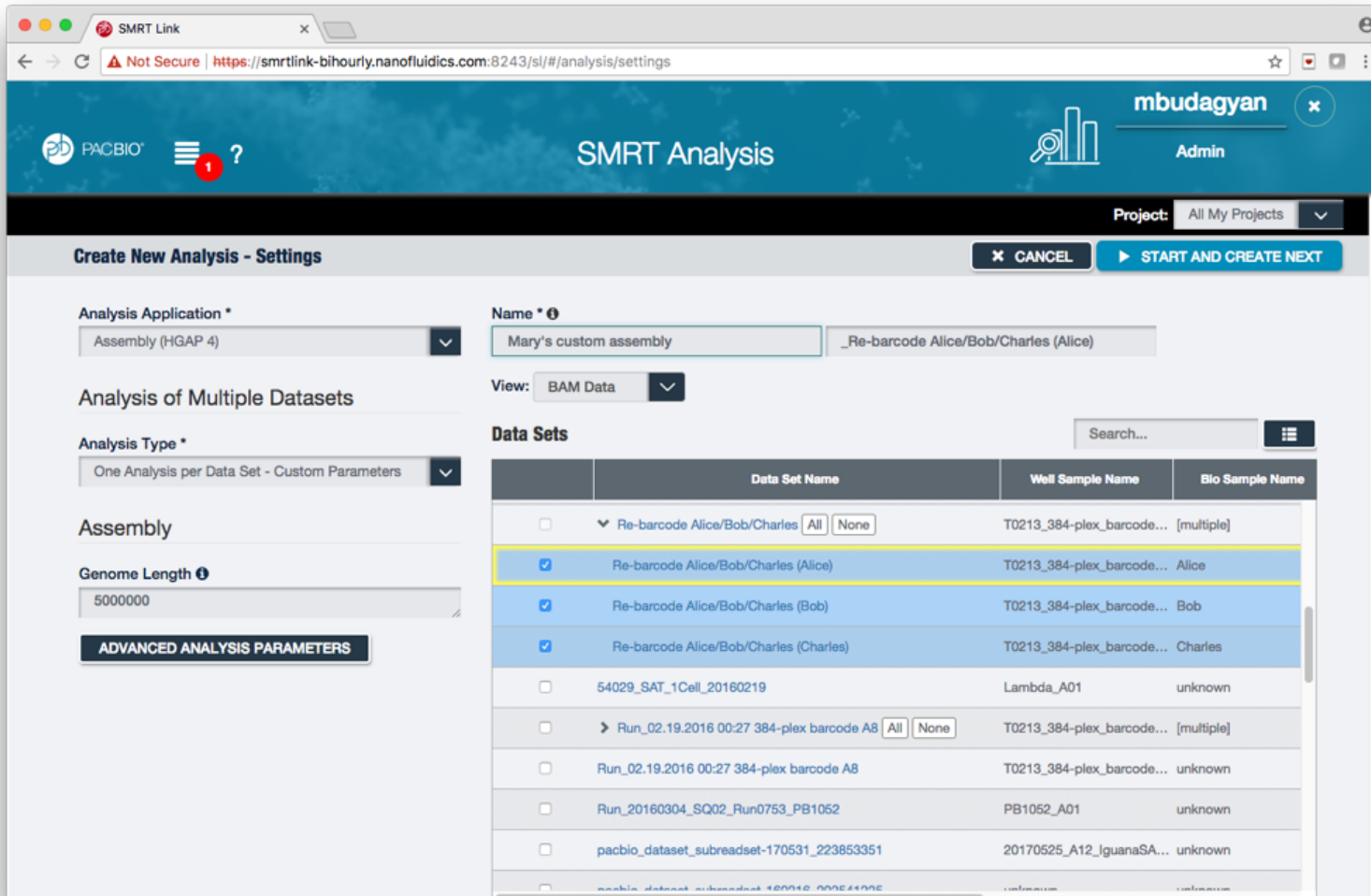
Name *

View: BAM Data

Data Sets Search...

	Data Set Name	Well Sample Name	Bio Sample Name
<input type="checkbox"/>	Re-barcode Alice/Bob/Charles All None	T0213_384-plex_barcode...	[multiple]
<input checked="" type="checkbox"/>	Re-barcode Alice/Bob/Charles (Alice)	T0213_384-plex_barcode...	Alice
<input checked="" type="checkbox"/>	Re-barcode Alice/Bob/Charles (Bob)	T0213_384-plex_barcode...	Bob
<input checked="" type="checkbox"/>	Re-barcode Alice/Bob/Charles (Charles)	T0213_384-plex_barcode...	Charles
<input type="checkbox"/>	54029_SAT_1Cell_20160219	Lambda_A01	unknown
<input type="checkbox"/>	Run_02.19.2016 00:27 384-plex barcode AB All None	T0213_384-plex_barcode...	[multiple]
<input type="checkbox"/>	Run_02.19.2016 00:27 384-plex barcode AB	T0213_384-plex_barcode...	unknown
<input type="checkbox"/>	Run_20160304_SQ02_Run0753_PB1052	PB1052_A01	unknown
<input type="checkbox"/>	pacbio_dataset_subreadset-170531_223853351	20170525_A12_IguanaSA...	unknown
<input type="checkbox"/>	pacbio_dataset_subreadset-160216_202541225	unknown	unknown

EDIT PARAMETERS FOR EACH DATASET, THEN LAUNCH



Create New Analysis - Settings

Analysis Application *
Assembly (HGAP 4)

Analysis of Multiple Datasets

Analysis Type *
One Analysis per Data Set - Custom Parameters

Assembly

Genome Length ⓘ
5000000

ADVANCED ANALYSIS PARAMETERS

Name * ⓘ
Mary's custom assembly _Re-barcode Alice/Bob/Charles (Alice)

View: BAM Data

Data Sets

	Data Set Name	Well Sample Name	Bio Sample Name
<input type="checkbox"/>	Re-barcode Alice/Bob/Charles <input type="button" value="All"/> <input type="button" value="None"/>	T0213_384-plex_barcode...	[multiple]
<input checked="" type="checkbox"/>	Re-barcode Alice/Bob/Charles (Alice)	T0213_384-plex_barcode...	Alice
<input checked="" type="checkbox"/>	Re-barcode Alice/Bob/Charles (Bob)	T0213_384-plex_barcode...	Bob
<input checked="" type="checkbox"/>	Re-barcode Alice/Bob/Charles (Charles)	T0213_384-plex_barcode...	Charles
<input type="checkbox"/>	54029_SAT_1Cell_20160219	Lambda_A01	unknown
<input type="checkbox"/>	Run_02.19.2016 00:27 384-plex barcode A8 <input type="button" value="All"/> <input type="button" value="None"/>	T0213_384-plex_barcode...	[multiple]
<input type="checkbox"/>	Run_02.19.2016 00:27 384-plex barcode A8	T0213_384-plex_barcode...	unknown
<input type="checkbox"/>	Run_20160304_SQ02_Run0753_PB1052	PB1052_A01	unknown
<input type="checkbox"/>	pacbio_dataset_subreadset-170531_223853351	20170525_A12_IguanaSA...	unknown

- SMRT Link highlights the current input dataset in yellow
- Autopopulates the dataset name to the (editable) analysis name



Structural Variation

Joint calling

MULTI-SAMPLE SELECTION FOR STRUCTURAL VARIATION

Create New Analysis - Settings Project: All My Projects ▼

Analysis Application *
Structural Variant Calling ▼

Analysis Name *
Multi-sample SV Demo

View: BAM Data ▼

Analysis of Multiple Datasets

Analysis Type *
One Analysis on All Data Sets ▼

Associated Inputs

Reference *
Hg19 ☰

Structural Variants

Minimum Length of Structural Variant (bp) ⓘ
50

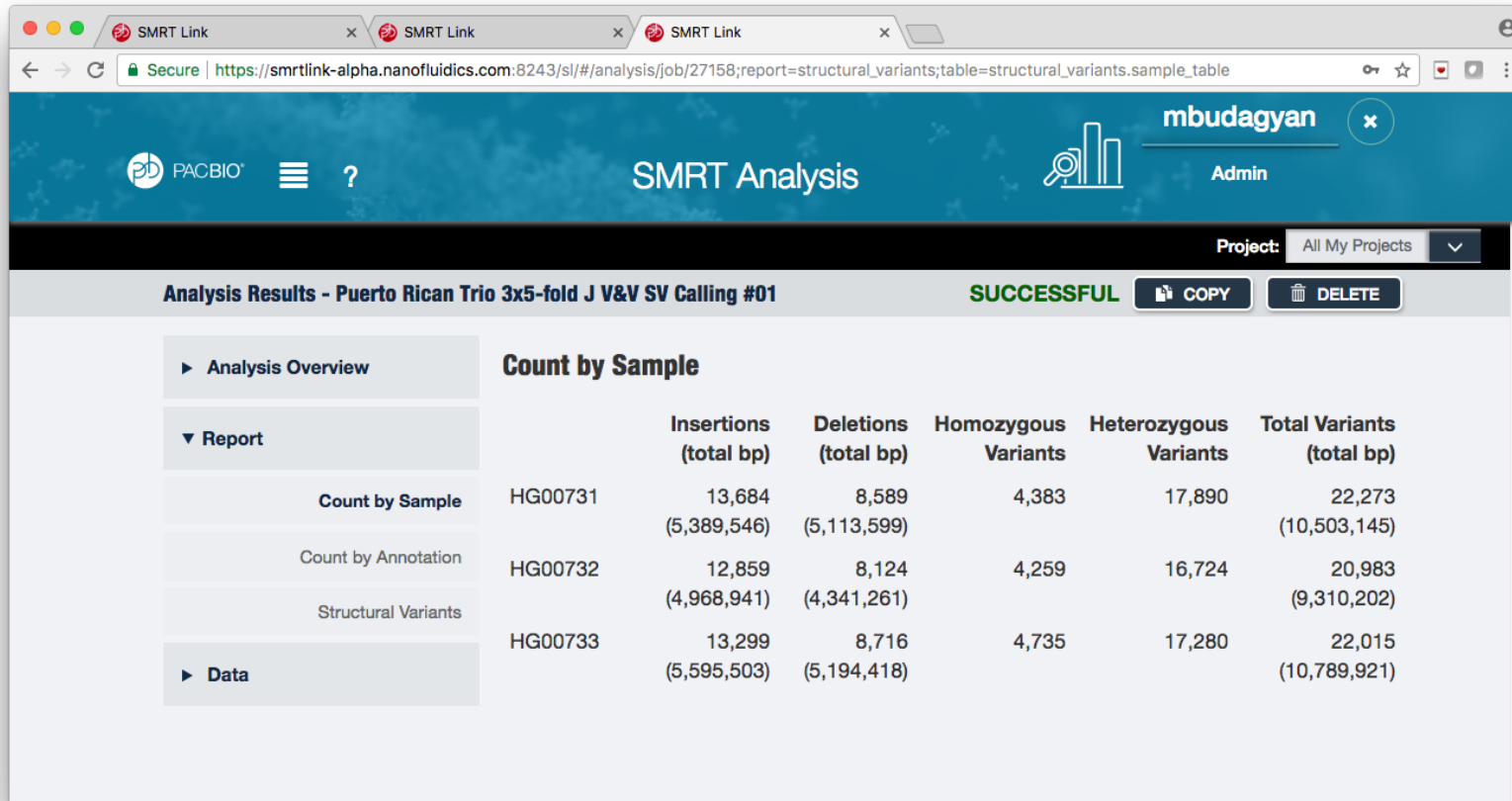
Minimum Reads That Support Variant (Count) ⓘ
2

Minimum Percentage of Reads That Support Variant (%) ⓘ
20

Data Sets hg00 ☰

	Data Set Name	Well Sample Name	Bio Sample Name	Barcode Name
<input checked="" type="checkbox"/>	HG00731-5fold J V&V	[multiple]	HG00731	
<input checked="" type="checkbox"/>	HG00732-5fold J V&V	[multiple]	HG00732	
<input checked="" type="checkbox"/>	HG00733-5fold J V&V	[multiple]	HG00733	
<input type="checkbox"/>	AW Merge Test	[multiple]	Testing	
<input type="checkbox"/>	HG00733_Jaguar_10hx3_20hx2	unknown	unknown	
<input type="checkbox"/>	HG00731 5-fold JV&V	unknown	unknown	
<input type="checkbox"/>	HG00733 10-fold JV&V	unknown	unknown	
<input type="checkbox"/>	HG00732 5-fold JV&V	unknown	unknown	
<input type="checkbox"/>	HG00733 5-fold JV&V	unknown	unknown	
<input type="checkbox"/>	.Jaguar SV VV 5cells 10hx3 20hx2	[multiple]	HG00733 .Jaguar 10hx3...	

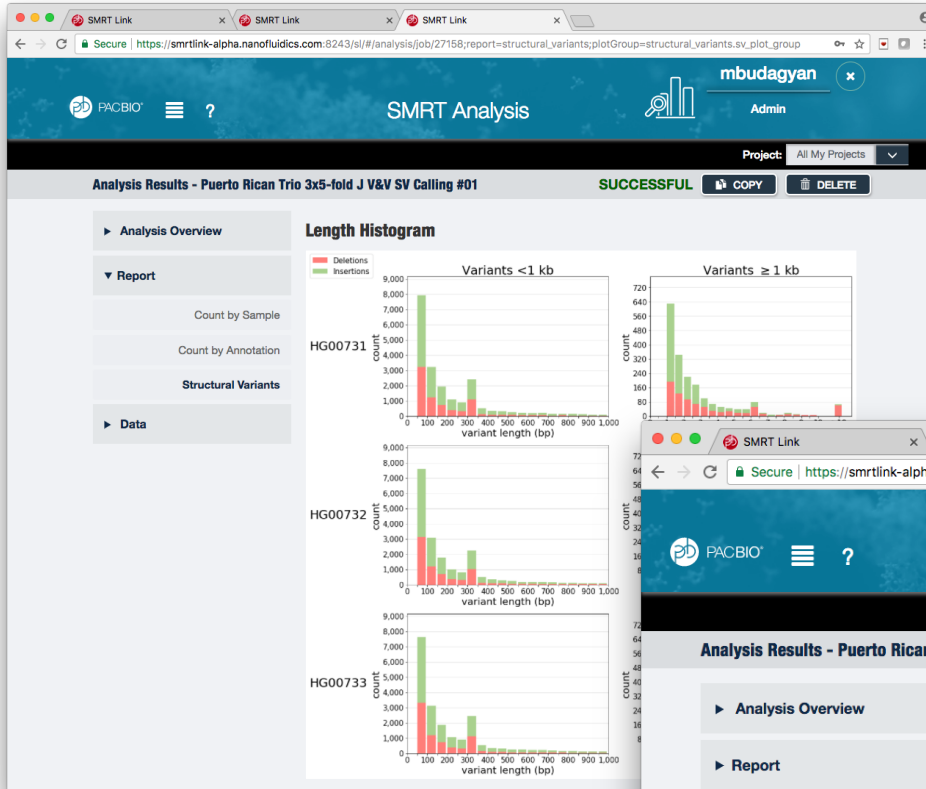
MULTI-SAMPLE SUPPORT



The screenshot shows the SMRT Analysis web interface. The browser address bar indicates the URL: https://smrtlink-alpha.nanofluidics.com:8243/sl/#/analysis/job/27158;report=structural_variants;table=structural_variants.sample_table. The user is logged in as 'mbudagyan' with 'Admin' privileges. The project is 'All My Projects'. The analysis results are for 'Puerto Rican Trio 3x5-fold J V&V SV Calling #01' and are 'SUCCESSFUL'. The interface includes a sidebar with navigation options: Analysis Overview, Report, and Data. The main content area displays a table titled 'Count by Sample' with columns for Insertions (total bp), Deletions (total bp), Homozygous Variants, Heterozygous Variants, and Total Variants (total bp). The table lists data for three samples: HG00731, HG00732, and HG00733.

		Insertions (total bp)	Deletions (total bp)	Homozygous Variants	Heterozygous Variants	Total Variants (total bp)	
Count by Sample	HG00731	13,684 (5,389,546)	8,589 (5,113,599)	4,383	17,890	22,273 (10,503,145)	
	Count by Annotation	HG00732	12,859 (4,968,941)	8,124 (4,341,261)	4,259	16,724	20,983 (9,310,202)
	Structural Variants	HG00733	13,299 (5,595,503)	8,716 (5,194,418)	4,735	17,280	22,015 (10,789,921)

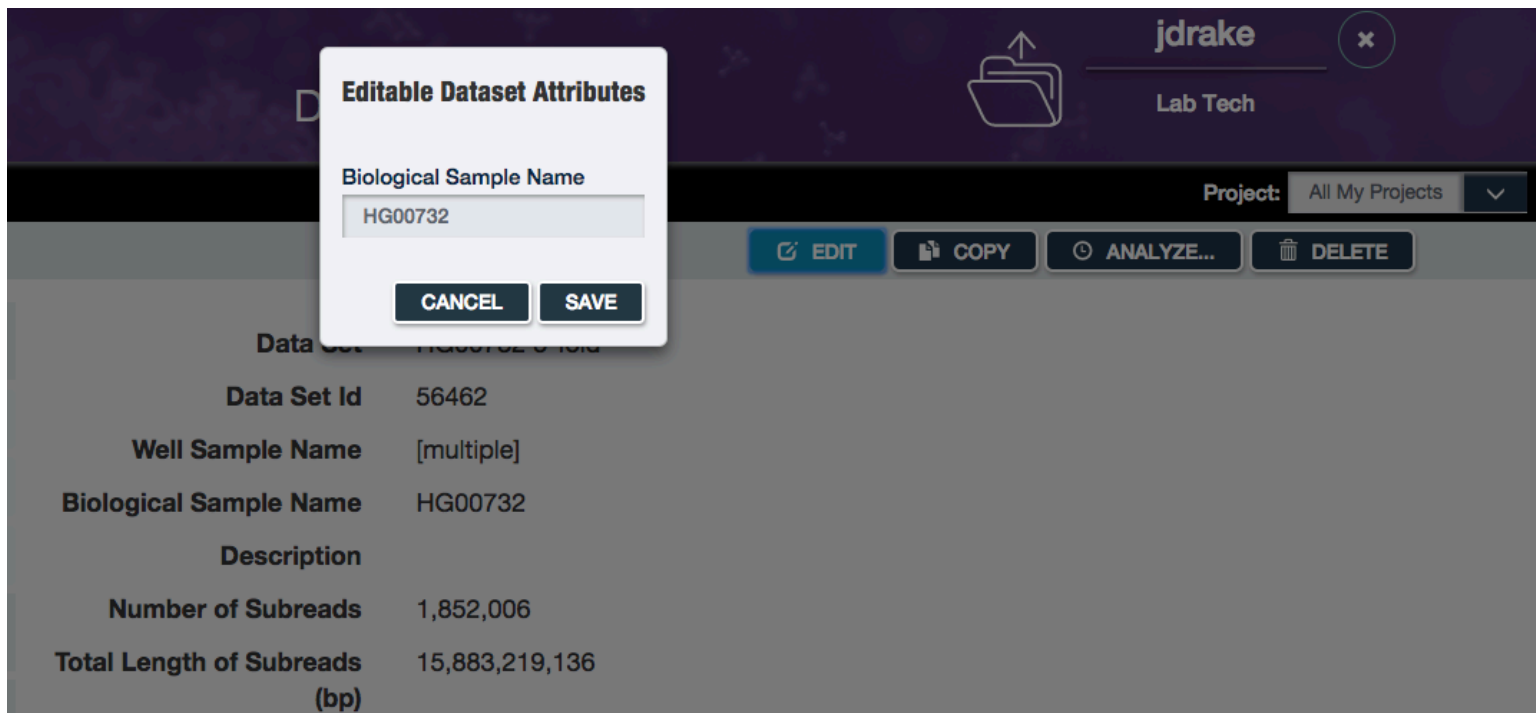
INFORMATION BROKEN UP BY SAMPLE



File Downloads

File	Size	Type
Aligned reads (HG00733)	9,433,744,160 bytes	bam
Aligned reads (HG00731)	9,277,861,072 bytes	bam
Aligned reads (HG00732)	9,495,217,252 bytes	bam
Analysis Log	59,225 bytes	log
Master Log	237,570 bytes	log
Structural variants	17,311,275 bytes	vcf
Structural variants	11,446,870 bytes	bed
Aligned reads	28,187,169,721 bytes	bam

BIOSAMPLE NAME



Editable Dataset Attributes

Biological Sample Name

CANCEL **SAVE**

Data Set HG00732

Data Set Id	56462
Well Sample Name	[multiple]
Biological Sample Name	HG00732
Description	
Number of Subreads	1,852,006
Total Length of Subreads (bp)	15,883,219,136

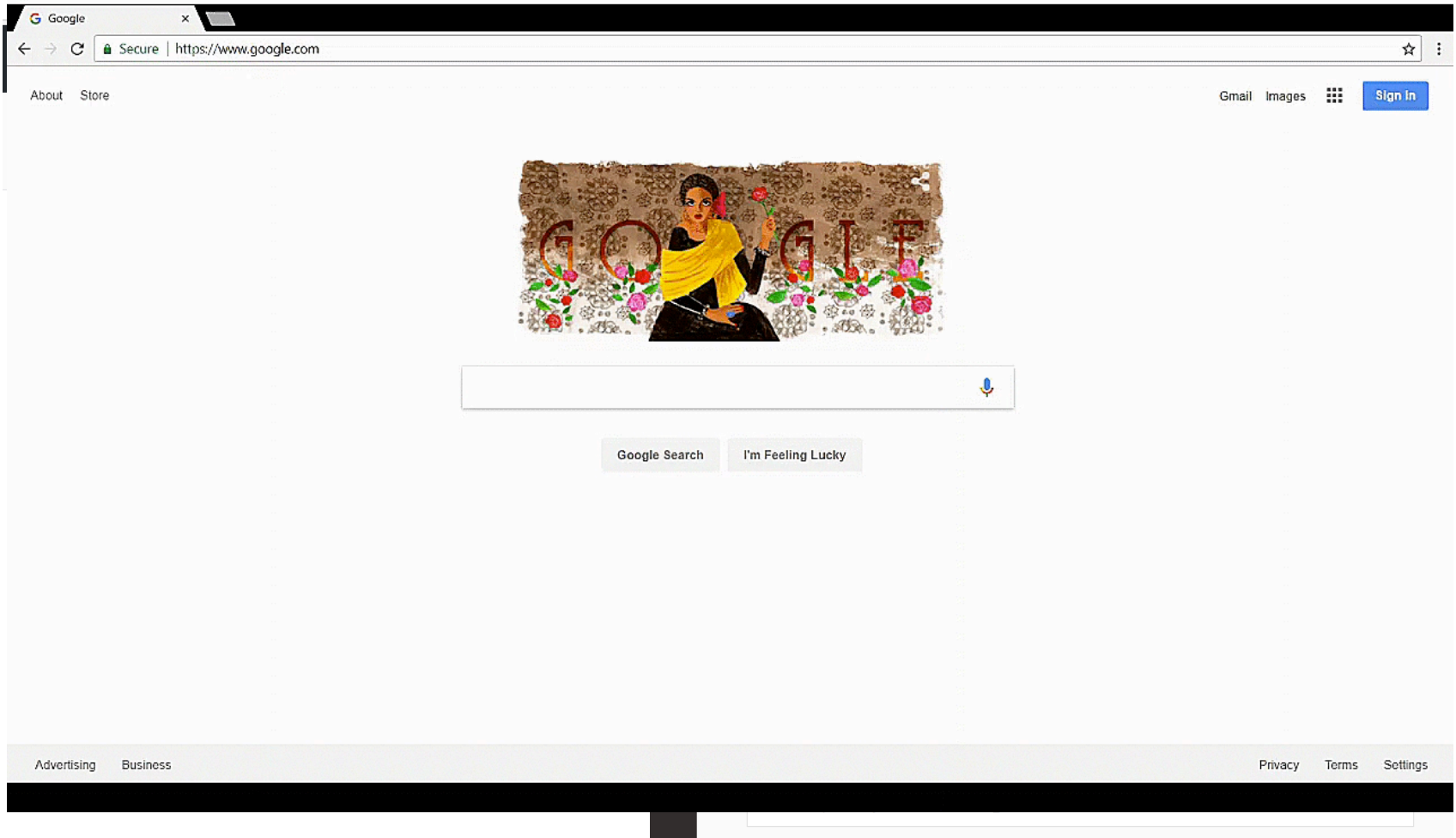
Project: All My Projects

EDIT **COPY** **ANALYZE...** **DELETE**



De Novo Assembly

BINARY RELEASE – FALCON/UNZIP



RUNNING UNZIP ON HGAP.4 OUTPUT

- HGAP 4 advanced parameter Save Output for Unzip, **off** by default
- Optionally retain the final set of `.las` files after overlapping raw reads, which can consume large amounts of disk space, especially for larger genomes
- Unnecessary when assembling haploid genomes, e.g., bacteria or when there is no intention to unzip

Advanced Analysis Parameters

30

Seed length cutoff ⓘ

-1

Save Output for Unzip ⓘ

ON OFF

Consensus

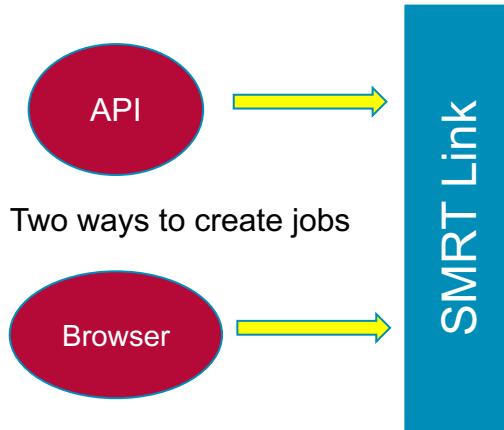
Algorithm ⓘ

best

Purpose ⓘ

variants

HGAP.4 TO UNZIP



Analysis Results - BDG_755_142_6_Sequel_Assembly SUCCESSFUL [SMRT View](#) [COPY](#) [DELETE](#)

▼ Analysis Overview	Analysis	BDG_755_142_6_Sequel_Assembly
	Analysis Id	2095
	Status	SUCCESSFUL: 132/132 total tasks completed.
	Created By	bgalvin
	Date Created	1/14/2018, 9:40:16 AM
	Date Updated	1/14/2018, 10:12:04 AM
► Polished Assembly	Application	Assembly (HGAP 4)
► Realignment to Draft Assembly	SMRT Link Version	5.1.0.21859
► Coverage	Inputs	BAM Data
► Preassembly	Path	/pbi/dept/secondary/siv/smrtlink/smrtlink-sms/smrtlink-release_5.1.0.14963/userdata/jobs_root/002/002095
► Data	Analysis Parameters	

```
# From the falcon code base
(venv) % python -m falcon_kit.mains.hgap4_adapt --help
usage: hgap4_adapt.py [-h] [--job-output-dir JOB_OUTPUT_DIR]
```

Given a full HGAP4 run, generate directories and symlinks to make it look like a pyeflow run.

optional arguments:

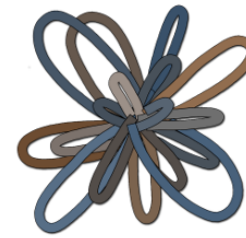
```
-h, --help          show this help message and exit
--job-output-dir JOB_OUTPUT_DIR
                    Directory of HGAP4 job_output. (A symlink or relative path is fine.) Task-dirs are under here in "tasks/" (de
```

Typically:

```
mkdir mydir/
cd mydir/
python -m falcon_kit.mains.hgap4_adapt --job-output-dir=../job_output/
```

GFA OUTPUT – ASSEMBLY DEBUGGING

Script available in both SMRT Tools and GitHub



Bandage

<https://github.com/rrwick>

Falcon
Assembly



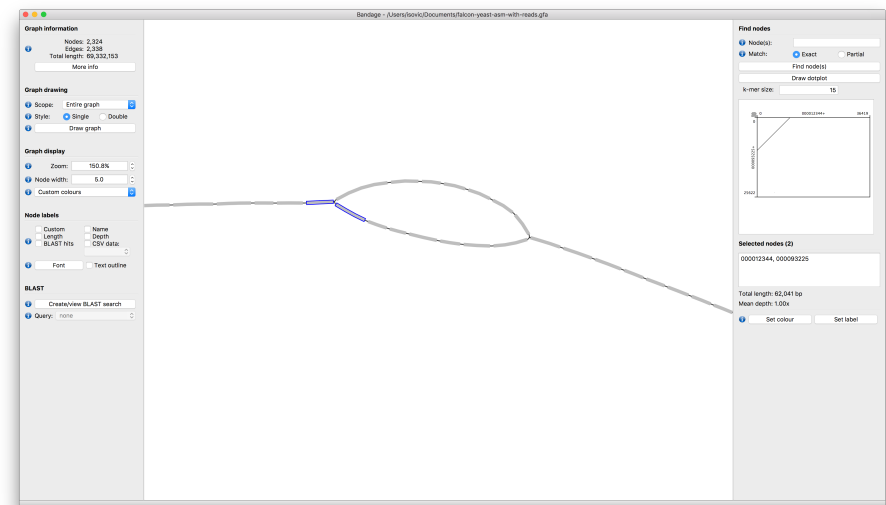
```
% python ./falcon_kit/mains/gen_gfa_v1.py -h
```

```
usage: gen_gfa_v1.py [-h] [--p-ctg-tiling-path P_CTG_TILING_PATH]
                  [--a-ctg-tiling-path A_CTG_TILING_PATH]
                  [--preads-fasta PREADS_FASTA] [--p-ctg-fasta P_CTG_FASTA]
                  [--a-ctg-fasta A_CTG_FASTA]
                  [--sg-edges-list SG_EDGES_LIST] [--utg-data UTG_DATA]
                  [--ctg-paths CTG_PATHS] [--add-string-graph]
                  [--write-reads] [--write-contigs] [--min-p-len MIN_P_LEN]
                  [--min-a-len MIN_A_LEN]
```

Generates GFA output (on stdout) from FALCON's assembly.



GFA
File





Usability Improvements

Data Management and SMRT Analysis

COPY AND RELAUNCH ANALYSIS

Project: All My Projects

Analysis Results - Demo SV Send Log Files FAILED **COPY** DELETE

Analysis Overview

Analysis Demo SV

Analysis Id 28612

Status FAILED: Task pbsvtools.tasks.align-18 FAILED in 295.08 sec.

Created By mbudagyan

Date Created 12/12/2017, 10:47:31 AM

Date Updated 12/12/2017, 10:55:59 AM

Application Structural Variant Calling

SMRT Link Version 5.1.0.20106

Inputs [BAM Data](#) [Reference](#)

Create New Analysis - Settings CANCEL START

Analysis Application * Structural Variant Calling

Analysis Name * Copy of Demo SV

Associated Inputs

Reference * Hg19

Structural Variants

Minimum Length of Structural Variant (bp) 50

Minimum Reads That Support Variant (Count) 2

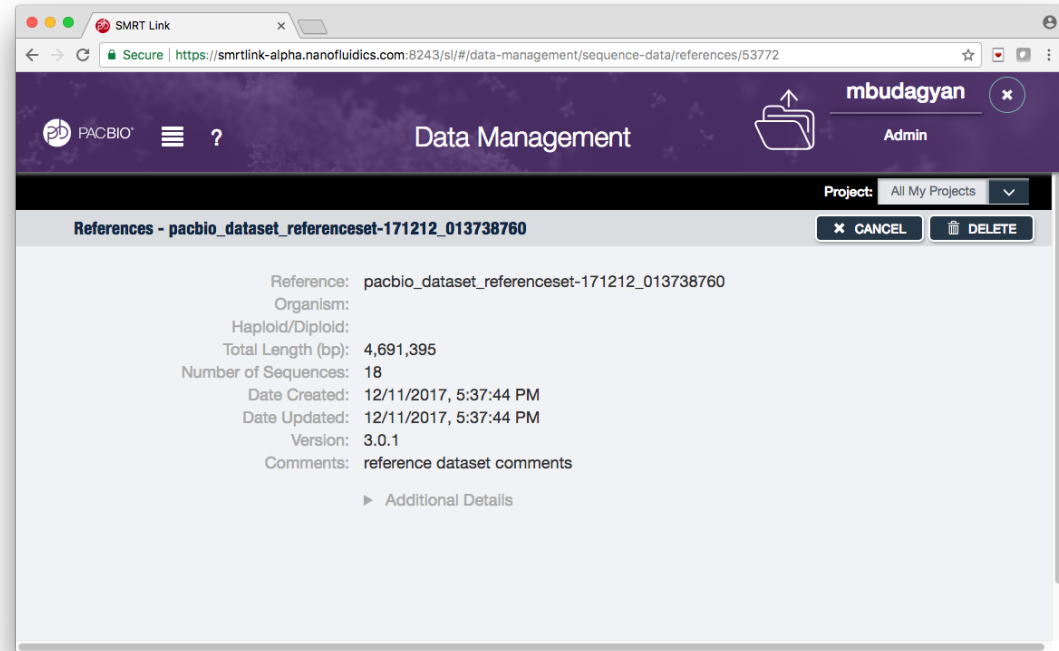
Minimum Percentage of Reads That Support Variant (%) 20

Data Sets

	Data Set Name	Well Sample Name	
<input type="checkbox"/>	zpmv-Cell1	zpmv	zpmv
<input type="checkbox"/>	BYU_Turneri-Cell5	BYU_Turneri	BYU_Turneri
<input checked="" type="checkbox"/>	HG00733 Sequel	unknown	unknown
<input type="checkbox"/>	3_Cell-1_Workflow_Diffusion.py-Cell8	3_Cell-1_Workflow_Diffusi...	3_Cell-1_Workflow_Diffusi...
<input type="checkbox"/>	15kb_Column Cleanup_5pM-Cell6	15kb_Column Cleanup_5pM	15kb_Column Cleanup_5pM 15
<input type="checkbox"/>	2_Cell-1_Workflow_Diffusion.py-Cell7	2_Cell-1_Workflow_Diffusi...	2_Cell-1_Workflow_Diffusi...
<input type="checkbox"/>	BYU_Turneri-Cell4	BYU_Turneri	BYU_Turneri
<input type="checkbox"/>	M1028-8pM-Cell4	M1028-8pM	M1028-8pM

DATA MANAGEMENT – REMOVING THE UNWANTED

- Deletes from the UI only
 - Collections
 - References
 - Barcode sets
- Helps reduce clutter
- UI responsiveness



CREATE ANALYSIS - NEW LAYOUT

Create New Analysis - Settings Project: All My Projects

CANCEL START

Analysis Application *

- Assembly (HGAP 4)
- Base Modification Detection
- Base Modification and Motif Analysis
- CCS Mapping
- Circular Consensus Sequences (CCS)
- Convert BAM to FASTX
- Demultiplex Barcodes
- Iso-Seq

Analysis Name *

View: BAM Data

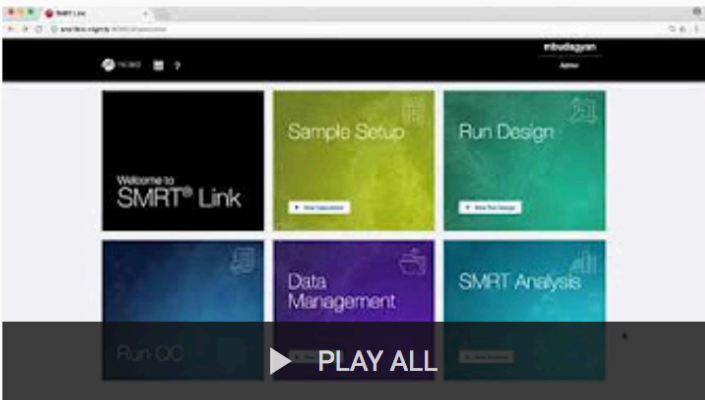
Data Sets Search...

	Data Set Name	Well Sample Name	
<input type="checkbox"/>	hdfsubreads	unknown	un
<input type="checkbox"/>	➤ Output-dataset-for-protractor-barcoding-test All None	BCS23 diffusion 70pM P6...	un
<input type="checkbox"/>	Auto-merged subreads @ 1513096418273	[multiple]	un
<input type="checkbox"/>	pacbio_dataset_subreadset-171114_215558341	10-10-2017_14hr_Lambd...	un
<input type="checkbox"/>	lambda/0007_tiny	Inst42267-040315-SAT-10...	un
<input type="checkbox"/>	pacbio_dataset_subreadset-170313_050639539	unknown	un
<input type="checkbox"/>	LVP2_D09-5253_BA008270-1st_4hrs_PkmidC525	"LVP2_D09-5253_BA0082...	un
<input type="checkbox"/>	➤ Pa_barcode_Alice/Bob/Charles All None	T0213_384_plex_barcode...	im

SMRT LINK REFERENCE MATERIALS



pacb.com > [Support](#) > [Software Downloads](#)


- SMRT Link Documentation
 - Release Notes
 - Installation Instructions
 - SMRT Link User Guide
 - Barcoding Overview
- Developer Documentation
 - SMRT Tools Reference Guide
 - SMRT Link APIs
- Training materials
 - SMRT Link and SMRT Analysis Tutorials



PacBio SMRT Link Training Series

9 videos • 324 views • Last updated on Dec 4, 2017

 **PacBio** [SUBSCRIBE](#)



www.pacb.com

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2018 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. FEMTO Pulse and Fragment Analyzer are trademarks of Advanced Analytical Technologies.

All other trademarks are the sole property of their respective owners.