

**Medical Sequencing Capabilities  
& Project Management  
2/1/2010**

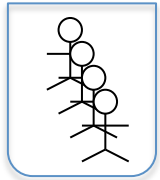
# Overview

- **Project Types (Menu of Capabilities)**
  - Small Designs: Few Genes & 1000's of Samples
  - Whole Exome: 100's – 1000's of Samples
  - Low Pass Shotgun: 4x genome coverage
- **Project Execution**
  - Concept/Design
  - Execution

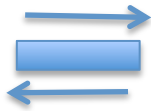
# few genes lots of samples

# all Genes Exome

1. Pool Samples

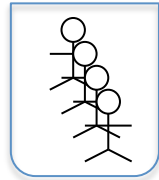


2. PCR

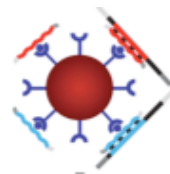


3. Sequence

1. Pool Samples

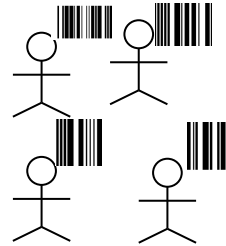


2. Selection

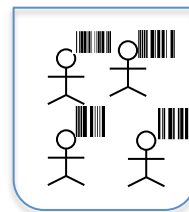


3. Sequence

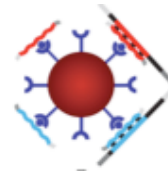
1. Barcode Samples



2. Pool

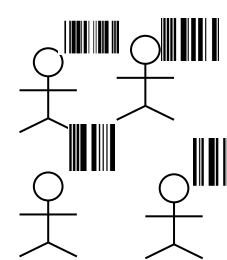


3. Selection

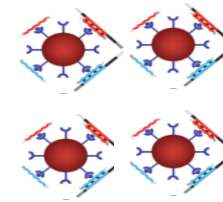


4. Sequence

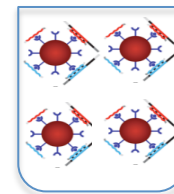
1. Barcode Samples



2. Selection



3. Pool

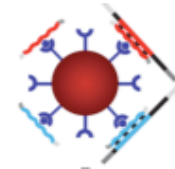


4. Sequence

1. Individual Sample



2. Selection



3. Sequence

- Effective targeting method
- Limit ~100 Genes (too many amplicons)
- Need follow up with genotyping

- Lower bound to target size (< 2Mb ~>50% off target)
- High duplication rate
- Need to follow up genotype

•High duplication rate

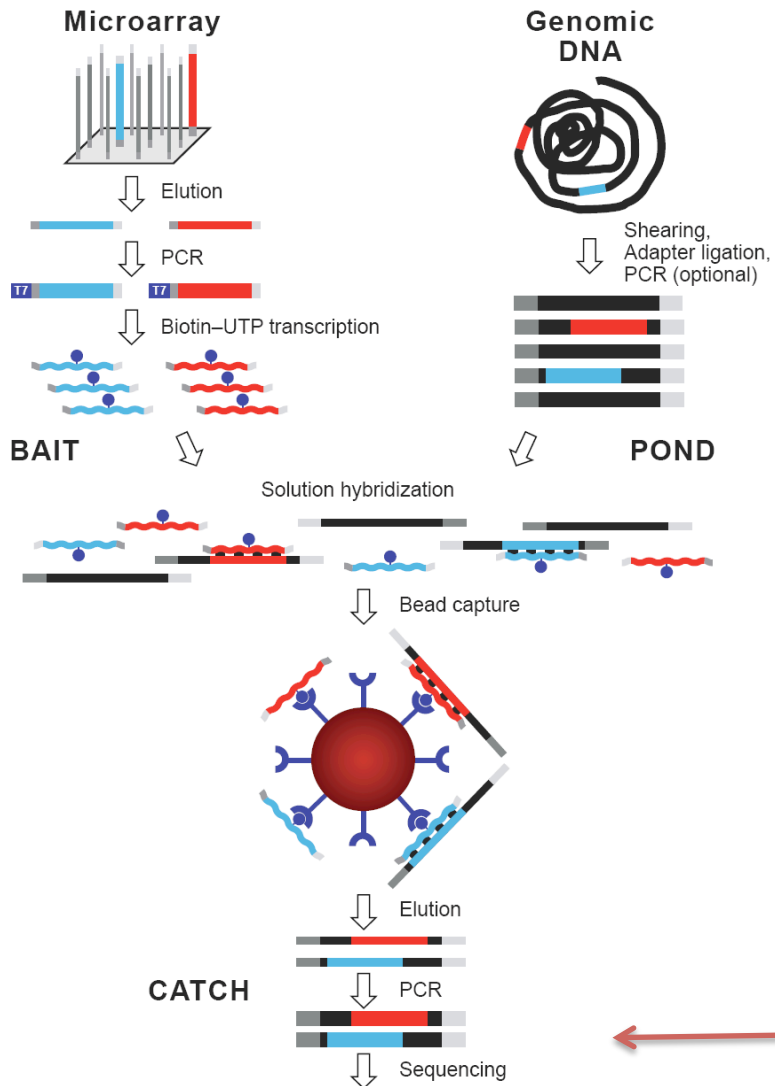
•Currently the best option for > 200 genes

- Efficient use of sequencing
- Current limitation is cost, takes 2 – 3 lanes/sample (\$15K)
- Projection \$5K within the year

Low Coverage Whole Genome Shotgun also possible

**small designs**

# Hybrid Selection Process

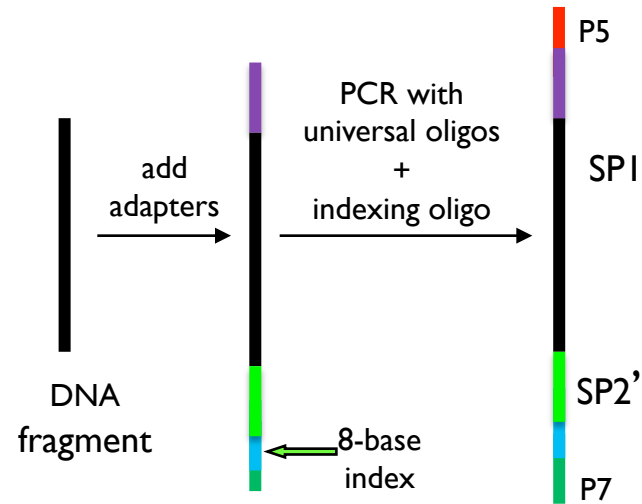


# With Barcoding...

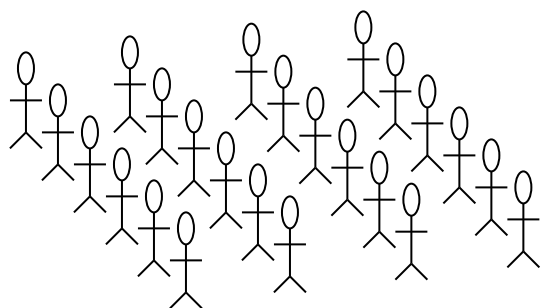
## External Adapter Indexing

V2.0 Adapter based Method, Released November 09

- 48 barcodes (8 base barcodes)
- Barcodes introduced at adapter ligation step
- 3<sup>rd</sup> Read indexing (68 bp + 8 bp index reads) or (101bp + 8 bp index reads)



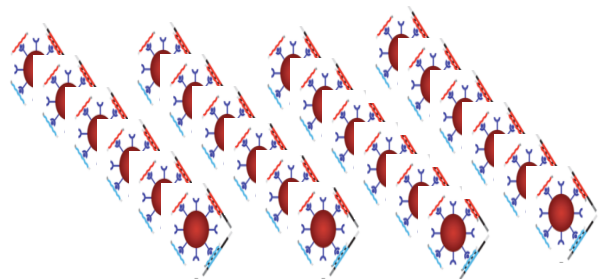
# Current Strategy for ~300 Genes x 1000+ Samples



24 Cases

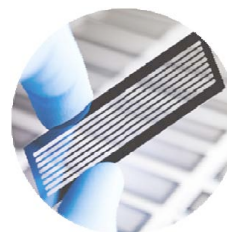


Indexing



Selection

Pooling



Sequencing  
2 Illumina Lanes

## Method:

Barcoded library hybrid capture

Illumina Sequencing

*76 bp paired end reads*

*24 sample multiplex pools*

*2 lanes per pool initially*

*Possible as yields increases to move to 1 lane/  
24 samples*

*Why not barcode 12 samples for 1 lane?*

*Barcode bases optimized for high quality  
sequencing in 3<sup>rd</sup> read technology.*

*Higher plexing results in higher quality  
data.*

# Proof of Principle – Hybrid Selection with Barcoding

- Selection Targets – FHS/Jackson Genes (216 genes + 1000 bp upstream of 5')
- Target: 798 kB
- Bait: 841 kB

## Pooling after selection 24 Samples in 4 lanes

	<b>Average</b>	<b>Min</b>	<b>Max</b>
<b>PF Reads</b>	4,358,231	1,106,094	7,597,618
<b>PF Bases - Unique</b>	242,840,674	64,537,828	413,189,607
<b>% Selected Bases</b>	28.0	22.5	32.5
<b>% Duplicate Reads</b>	5.60	1.78	9.30
<b>Mean Target Coverage</b>	63.6	19.4	100.3
<b>Fold 80 Penalty</b>	2.8	2.6	3.2
<b>% of Target Bases &gt; 20x</b>	79.3	43.0	89.4
<b># SNPs</b>	479.3	254.0	592.0
<b>% SNPs in dbSNP</b>	82.5	77.4	91.3

**With 12 sample plexing: Estimate ~ 30x mean target coverage  
90% of bases > 10x**

# Some Numbers for Project Planning (small designs)

CURRENT		
Number of Genes	250	
Target Size	800,000	
# of samples per pool	50	
coverage required	20	
sequence required	800,000,000	
unevenness penalty	5	
total sequence required	4,000,000,000	
yield per lane	2,500,000,000	2.5 G PF/lane
% on target	1,250,000,000	(50%).
% non duplicate	1,062,500,000	(15%).
lanes for 50 samples	3.8	
sample plexing possible	13/lane	



Current plex = 12

TARGET		
Number of Genes	250	
Target Size	800,000	
# of samples per pool	50	
coverage required	20	
sequence required	800,000,000	
unevenness penalty	5	
total sequence required	4,000,000,000	
yield per lane	3,000,000,000	3G PF/lane
% on target	2,400,000,000	(80%).
% non duplicate	2,040,000,000	(15%).
lanes for 50 samples	2.0	
sample plexing possible	25	



Target plex = 24

# Whole Exome Capabilities

# Whole Exome Bait Designs

---

	Exome Panel 1	Exome Panel 2
<b># Genes</b>	15,994	18,560
<b># Targets</b>	164,688	188,260
<b>Target Territory</b>	28.6 Mb	32.7 Mb
<b># Baits</b>	316 K	-
<b>Bait Length</b>	120b	120b
<b>Bait Territory</b>	37.6 Mb	44.9 Mb
<b>Accessible Target*</b>	26.6 Mb	-

~500 samples processed

In Production  
2/1/10

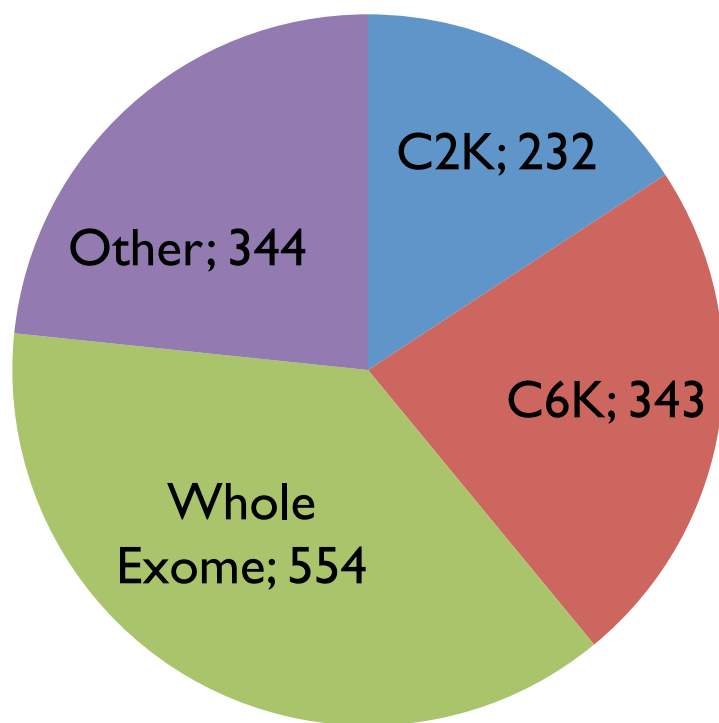


Agilent Technologies



# Whole Exome Production Experience

---

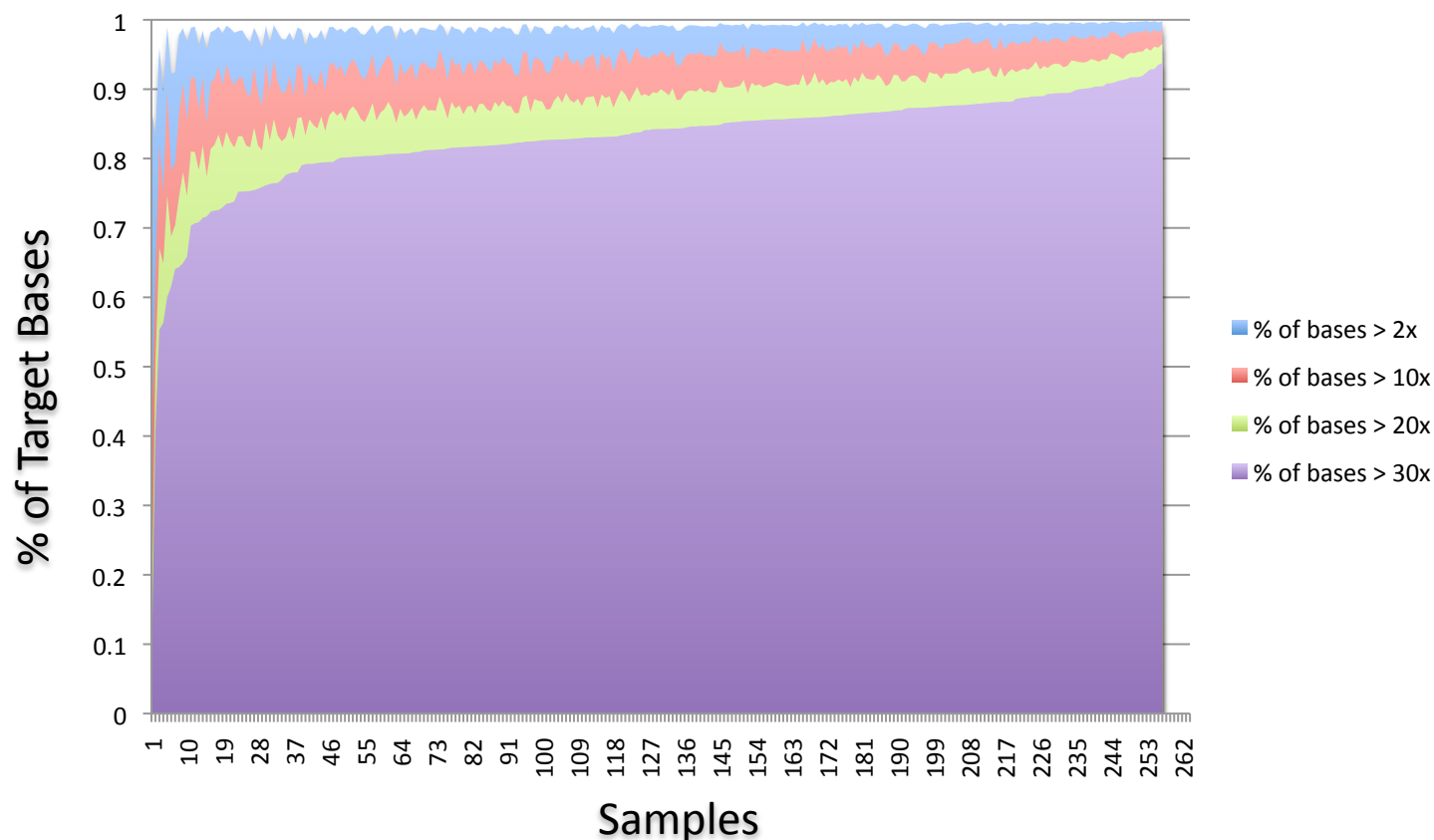


All Hybrid Selection  
Production Since July 09

1,473 Libraries!

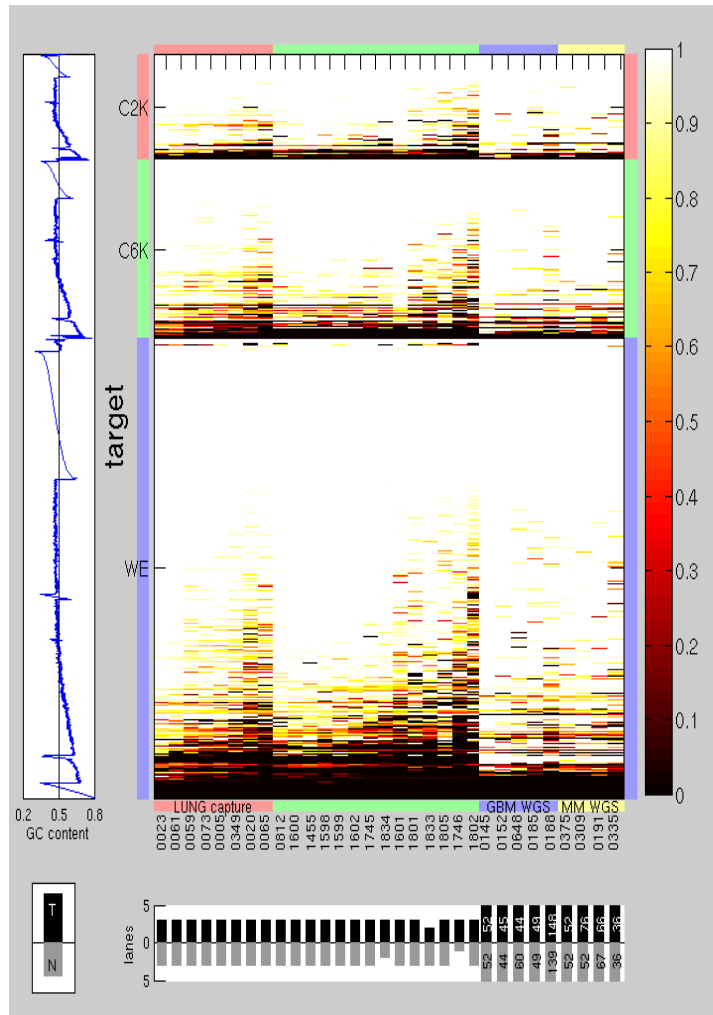
2/3 whole exome

# Whole Exome Production Experience



- 554 Whole Exome libraries generated to date (Exome Panel I)
- Most are 3 Lanes/sample
- 247/258 analyzed samples have > 80% of alignable target bases covered
- Standard analysis results 90 % of SNPs called are present in dbSNP
  - Mean # of SNPs called = 17,000 (in alignable/reachable target 25 Mb)

# Whole Exome Coverage – Production Samples



## Application: Cancer Somatic Mutation Discovery

- Coverage assessed on Tumor & Normal Paired Samples
- Covered Base = 8 reads in Normal, 14 in Tumor
- GC bias evident in whole exome & whole genome
- Important to measure exome coverage by alignability

Whole Exome

WGS

# Coverage of Genes (n=15,994)

## Percentage of Genes

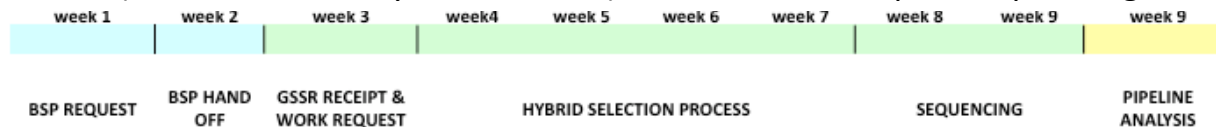
With at least

	@10X	@20X	@30X
>70% of bases	91%	85%	77%
>80% of bases	86%	76%	66%
>90% of bases	75%	62%	47%

# Frequently Asked Questions about Targeted Sequencing Projects

- **How much DNA do I need?**  
3ug for sequencing + ~100ng for fingerprinting  
All samples need QC, Quantification(pico), and Proof of IRB approval in BSP
- **Can I use WGA?**  
Yes, but WGA product needs column clean up (usual WGA caveats apply)
- **How long does it take to design & receive targeting oligos?**  
~1 month (1 week to refine targets & design, 3 weeks for agilent to synthesize oligos & prep RNA bait)
- **What is the current library construction & sequencing capacity?**  
4 Plates (376 samples) per week, Targeting scale up to 700 in Q2, 1700 in Q4

- **How long will the sequencing production take?**  
~ 6 weeks (4 weeks for library construction) + 2 weeks for 76bp PE sequencing & analysis

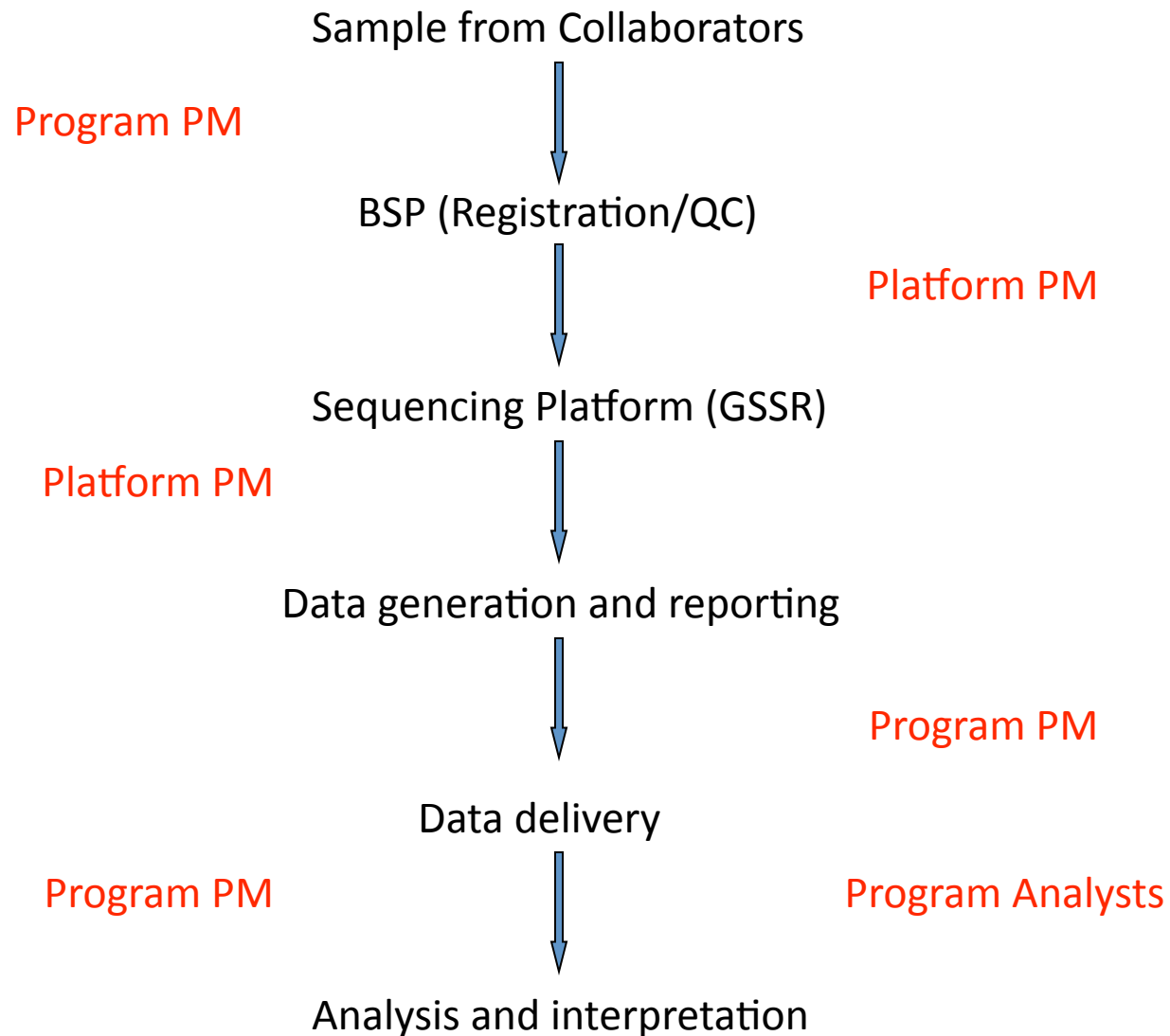


- **What is the current success rate?**  
Difficult to measure as different projects have different coverage goals but early batches are encouraging.
- **How can I get a project started?**  
More on this!

# Initiating & Executing a Project

- Capacity is tight, but will improve as scale increases
- Contact the project management group as early as possible:
  - Noel Burtt: [burttn@broadinstitute.org](mailto:burttn@broadinstitute.org)
  - Carrie Sougnez [carrie@broadinstitute.org](mailto:carrie@broadinstitute.org)
- Get samples into BSP & QCed
  - Fingerprinting is strongly recommended for concordance matching post analysis
- Project managers will take it from here
  - Scheduling of project in the capacity queue and providing a timeline
  - Filling out required paperwork for platform work
  - Overseeing execution
- Deliverables:
  - Alignment files (BAM)
  - SRA Data (for some projects submitted to DCC/NCBI)
  - Sequencing quality reports (squid)

# Workflow and hand offs:



# Project Managers (people who can help!)

## Medical and Population Genetics

Noel Burtt – Program Manager

Candace Guiducci

*Diabetes, Metabolic Diseases, Rheumatoid  
Arthritis*

Christine Stevens

*Autism, MCKDI Kidney*

Namrata Gupta

*HIV*

Jennifer Moran

*Schizophrenia, Bipolar Disorder*

Deb Farlow

*NHLBI Cohorts GO Exome Project*

## Cancer/NHGRI Initiatives

Carrie Sougnez, Erica Shefler

*TCGA, TSP, 1000 Genomes, NHGRI Medical  
Sequencing*

Rob Onofrio

*Cancer (GAP)*

## Sequencing Platform Managers

Jane Wilkinson

Lauren Ambrogio

Mugdha Velankar

Email: [medcancerpm@broadinstitute.org](mailto:medcancerpm@broadinstitute.org)

end