Post Assembly NGS Analysis

A Core Perspective

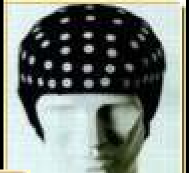
Brent G. Richter

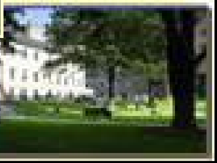
Director, ERIS

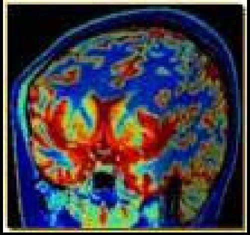
PARTNERS HEALTHCARE | FOUNDED BY BRIGHAM AND WOMEN'S HOSPITAL AND MASSACHUSETTS GENERAL HOSPITAL

PARTNERS HEALTHCARE

FOUNDED BY BRIGHAM AND WOMEN'S HOSPITAL
AND MASSACHUSETTS GENERAL HOSPITAL





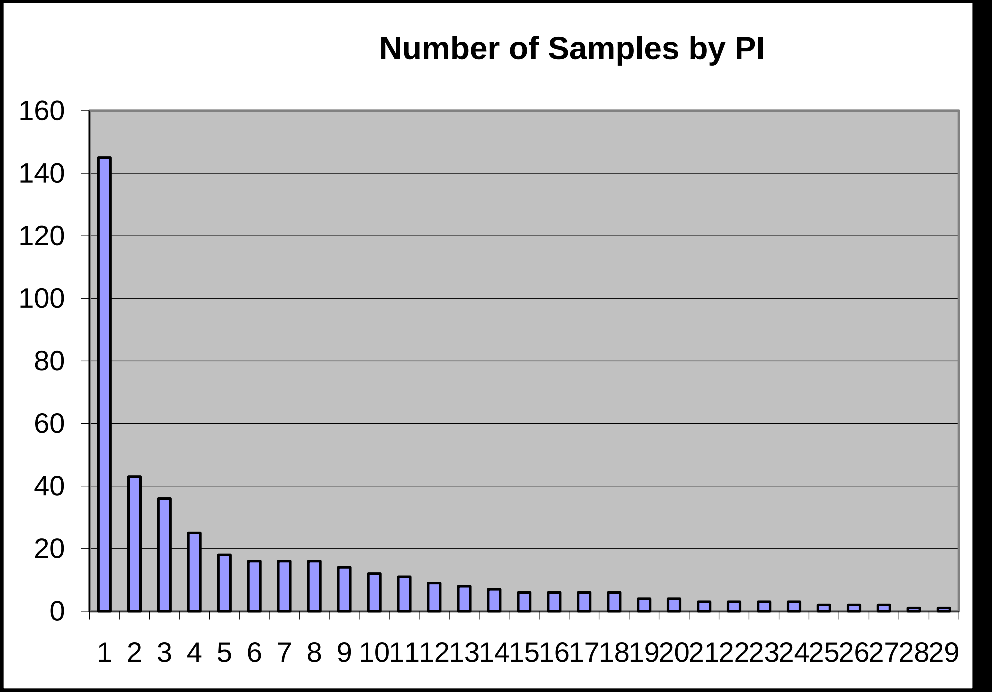


NGS at PHS

- 4 Solexa Instruments deployed at PHS
  - 2 cores/geographic locations: 3 PCPGM, 1 MGH
  - 1 Additional machine in 6 months
- (2 Helicos Systems in testing)
- 3.0TB per instrument per week
- Pre-analysis pipeline developed and maintained by ERIS
  - Within HPC environment
  - Test new base callers/assemblers

Distribution of Service

- One Bioinformaticist
  - Able to care/feed pipeline, deliver alignments
  - Investigators primarily perform further analysis
- Since Jan 09
  - 283 alignments
  - 145 raw data



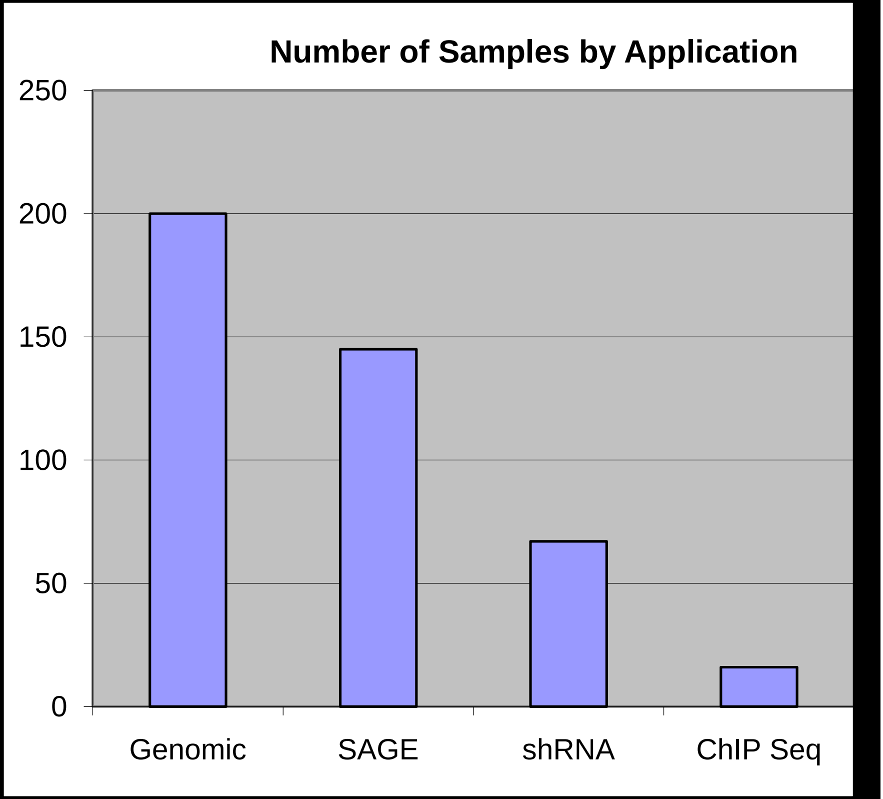
Applications

Category	**Examples of applications**
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Adopted from Shendure & Ji, Nat Biotech 2008

Experimental Applications

- Targeted genomic re-sequencing
  - Reference alignment
- Small RNA sequencing
- Serial Analysis of Gene Expression (SAGE)
- Chip-Seq
- Currently multi-plexing for better economy



Partial list of Analysis Tools

Program	**Categories**	**Program**	**Categories**
Cross_match	Alignment	Edena	Assembly
ELAND	Alignment	Euler-SR	Assembly
Exonerate	Alignment	SHARCGS	Assembly
MAQ	Alignment and variant detection	SHRAP	Assembly
Mosaik	Alignment	SSAKE	Assembly
RMAP	Alignment	vCAKE	Assembly
SHRiMP	Alignment	velvet	Assembly
SOAP	Alignment	PyroBayes	Base caller
SSAHA2	Alignment	PbShort	variant detection
SXOligoSearch	Alignment	ssahaSNP	variant detection
ALLPATHS	Assembly		

From: http://seqanswers.com/forums/showthread.php?t=43
And http://www.sanger.ac.uk/Users/lh3/seq-nt.html

Current Analysis Tools

Program	**Purpose**
Cross_match (David Gordon)	Alignment
ELAND (Illumina)	Alignment
MAQ (Sanger)	Alignment and Variant detection
VAAL (Broad MIT)	Alignment and Variant detection

Commercial Program	**Purpose**
GenomeQuest	Variant detection, de-novo assembly, tag counts
Genomatix	Variant detection, de-novo assembly, tag counts, and transcription factor analysis
CLCBio	-Same-

Challenges

- Alignment of short reads
- Lack of standards for cross-comparisons
- Choosing the right algorithms for sequencing applications

Challenges—alignment of reads

- Blast/Blat do not work well
- Commercial and open-source algorithms
  - Strength/weakness :: speed/quality
- Evaluation of each algorithm for specific purpose

Challenge—lack of standards

- Comparing quality of bases across platforms/versions/alignments
- Accepted practice: convert all qualities to phred-like scores

Challenge—the right algorithm

- Standard comparisons
- One algorithm, specific application
- Best practice: experimentation