Analysis of Large Datasets

ISMB 2010 – Bioinformatics Cores Workshop

Boston, Mass

July 11, 2010

Overview

- 2 Presentations
  - Best practices working with large datasets – *Dawei Lin*
  - Integrating data and meta-analysis of publically available expression array data, looking for shared pathways – *Vared Caspi*
- Moderated question/discussion session
  - Methods for QC
  - Triage of source data
  - Open source software
  - Analysis methods
  - Developing software encapsulating new analysis methods

Data “Tsunami”

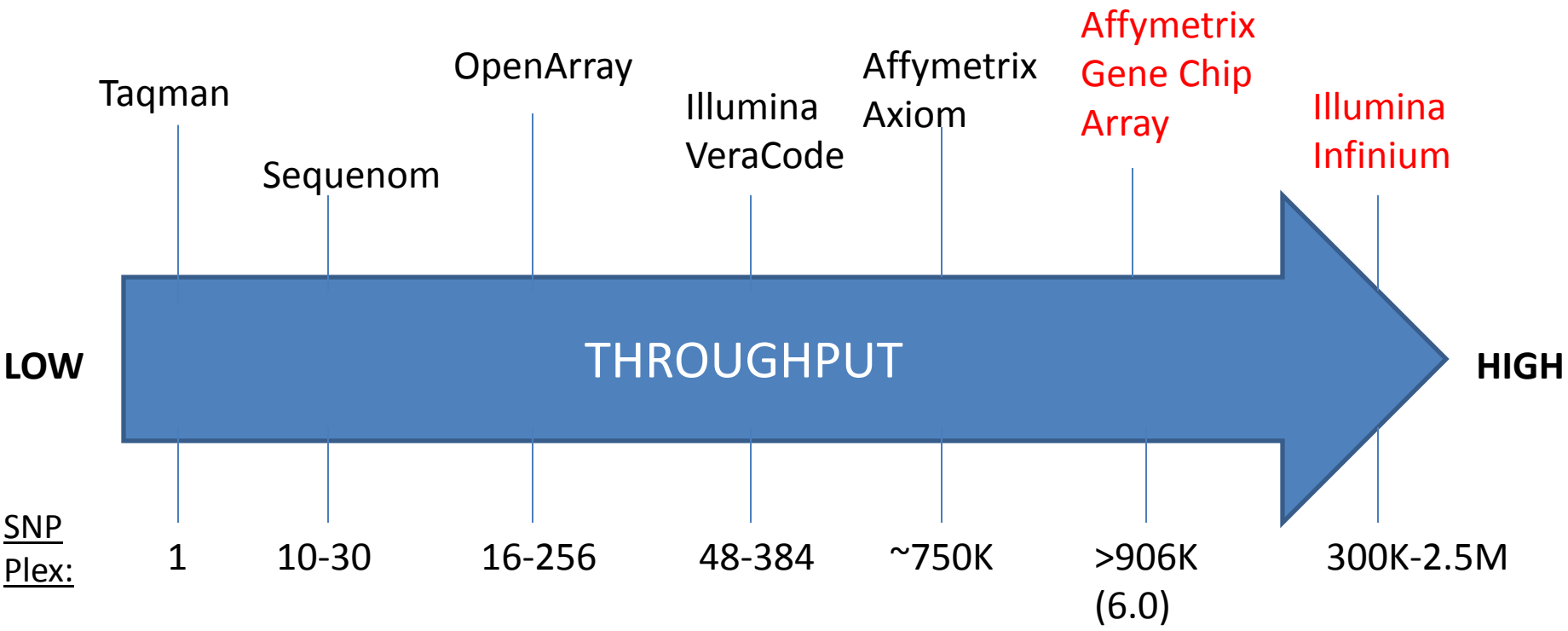
Hardware for Sequencing

Hardware for Genotyping

ay Genotyping System, see the feature article,
com/taqmanopenarray

Figure 1. TaqMan® OpenArray™ Genotyping System.

Instrument Throughput



SNP Plex:	1	10-30	16-256	48-384	~750K	>906K (6.0)	300K-2.5M
# Samples Per "Run":	384	384	144-12	96	96	1-5*	96-16*

Bioinformatic Challenges Abound

- How/What to store?
- How to process?
- How to analyze?
- How to assess quality?
  - What is quality?
  - How to QC?

Next Generation Sequence Data

Next-Generation Sequencing Statistics

Vendor:	Roche			Illumina			ABI		
Technology:	454			Solexa GA			SOLiD		
Platform:	GS20	FLX	Ti	I	II	IIx	1	2	3
Reads: (M)	0.5	0.5	1.25	28	100	250	40	115	320
Fragment									
Read length:	100	200	400	35	50	100	25	35	50
Run time: (d)	0.25	0.3	0.4	3	3	5	6	5	8
Yield: (Gb)	0.05	0.1	0.5	1	5	25	1	4	16
Rate: (Gb/d)	0.2	0.33	1.25	0.33	1.67	5	0.34	1.6	2
Images: (TB)	0.01	0.01	0.03	0.5	1.1	2.8	1.8	2.5	1.9
PA Disk: (GB)	3	3	15	175	300	300	300	750	1200
PA CPU: (hr)	10	140	220	100	70	NA	NA	NA	NA
SRA: (GB)	0.5	1	4	30	50	2.5	100	140	600
Paired-end									
Read length:		200	400	2×35	2×50	2×100	2×25	2×35	2×50
Insert: (kb)		3.5	3.5	0.2	0.2	0.2	3	3	3
Run time: (d)		0.3	0.4	6	10	10	12	10	16
Yield: (Gb)		0.1	0.5	2	9	50	2	8	32
Rate: (Gb/d)		0.33	1.25	0.33	1.67	5	0.34	1.6	2
Images: (TB)		0.01	0.03	1	2.2	5.6	3.6	5	3.8
PA Disk: (GB)		3	15	350	500	550	600	1500	2400
PA CPU: (hr)		140	220	160	120	NA	NA	NA	NA
SRA: (GB)		1	4	60	100	3.5	200	280	1200

Do you engage in data triage?

From David Dooling at Politigenomics