BIOINFORMATICS & HPC CORE

PRINCESS MARGARET CANCER CENTRE, UHN

Who are we and what do we do?

Natalie Stickle & Zhibin Lu

OUTLINE

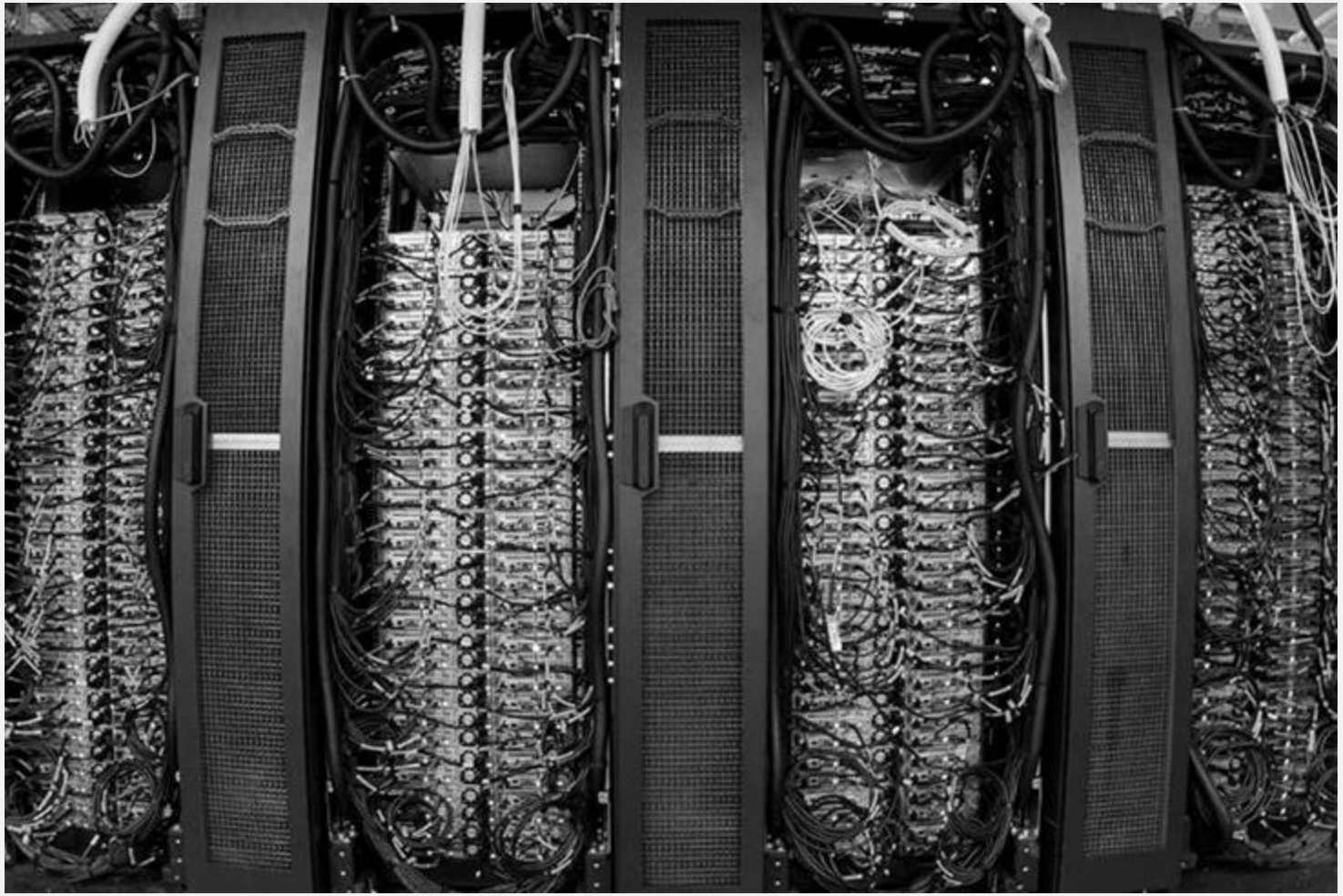
- What is Bioinformatics?
- What is HPC?
- The data problem.
- Bioinformatics & HPC Core mandate
- Internal HPC capabilities
- HPC4Health
- Core Services
  - Bioinformatic Analyses
  - Single Cell Genomics
  - Web sites, wikis and CMS
- Alignment within Canada and beyond

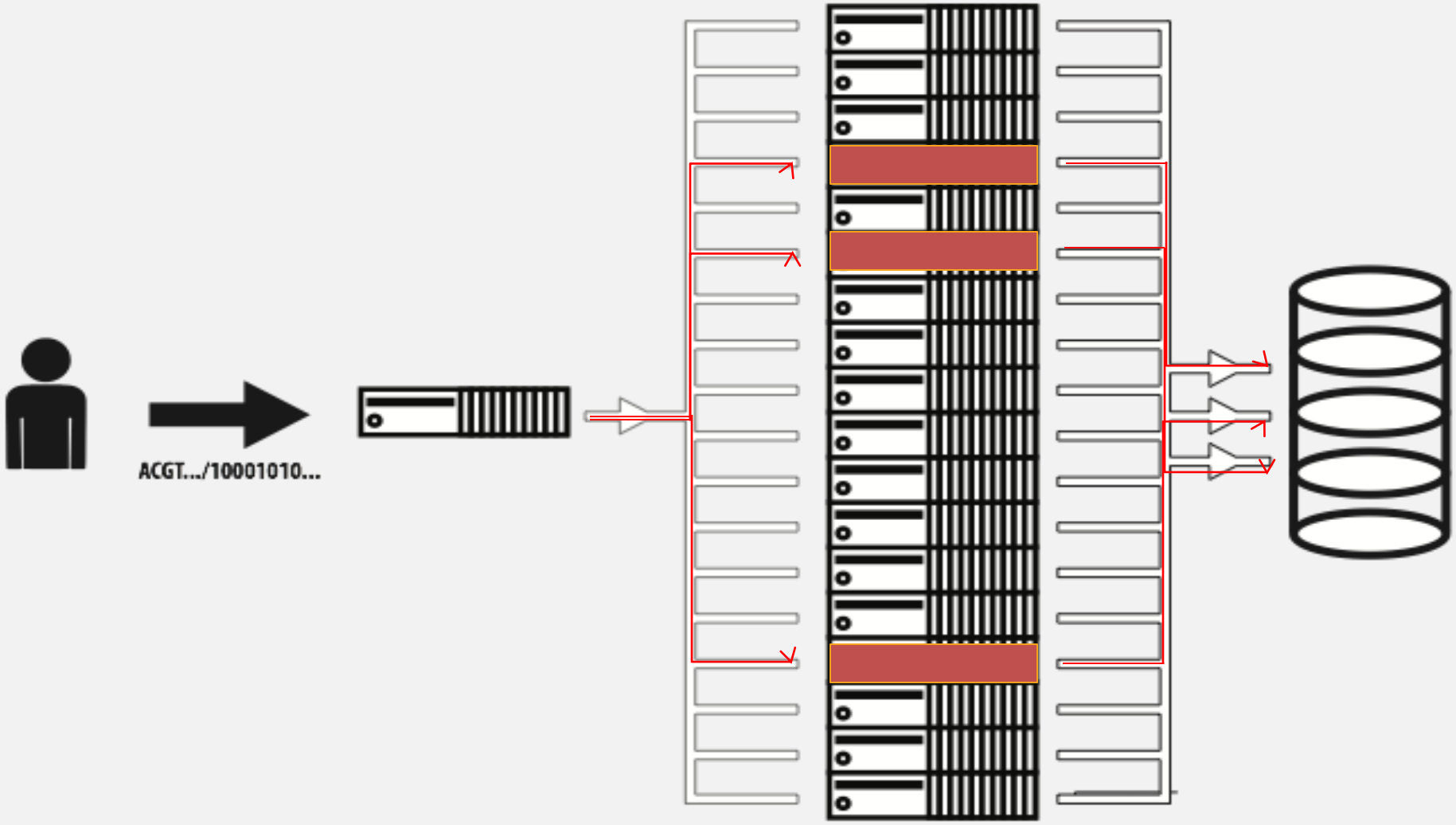
WHAT IS BIOINFORMATICS?

- A "catch-all" phrase that means different things to different people
- Intersection of **computer science**, **mathematics** and **statistics** - application of these to problems in areas of basic biology such as **genomics**, **proteomics**, molecular biology, etc.
- Often employs state-of-the-art techniques in **artificial intelligence** and **machine learning** to create meaning and extract useful information from extremely large datasets
- Usually needs access to **High-Performance Computing (HPC)** resources due of size of data to be analyzed and complexity of the processing

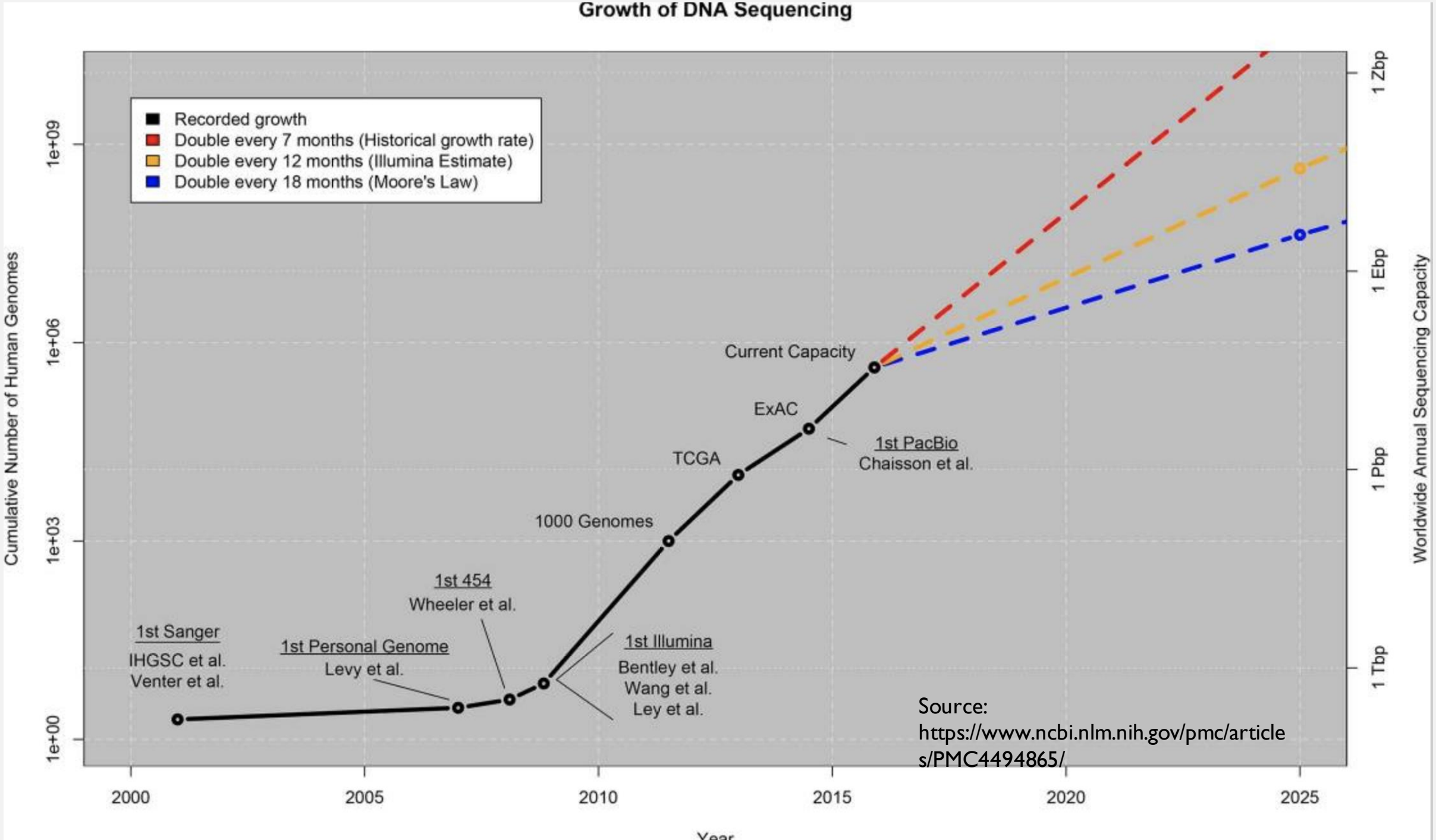
A SHORT INTRODUCTION TO HPC

- Multiple servers connected together and integrated (called a cluster)
  - Each server in the cluster is called a node
- High speed connections
- Storage is optimized for fast I/O
  - eg. 80 Gb/s server vs. MAX 6 Gb/s desktop
- Scheduler software responsible for access to nodes



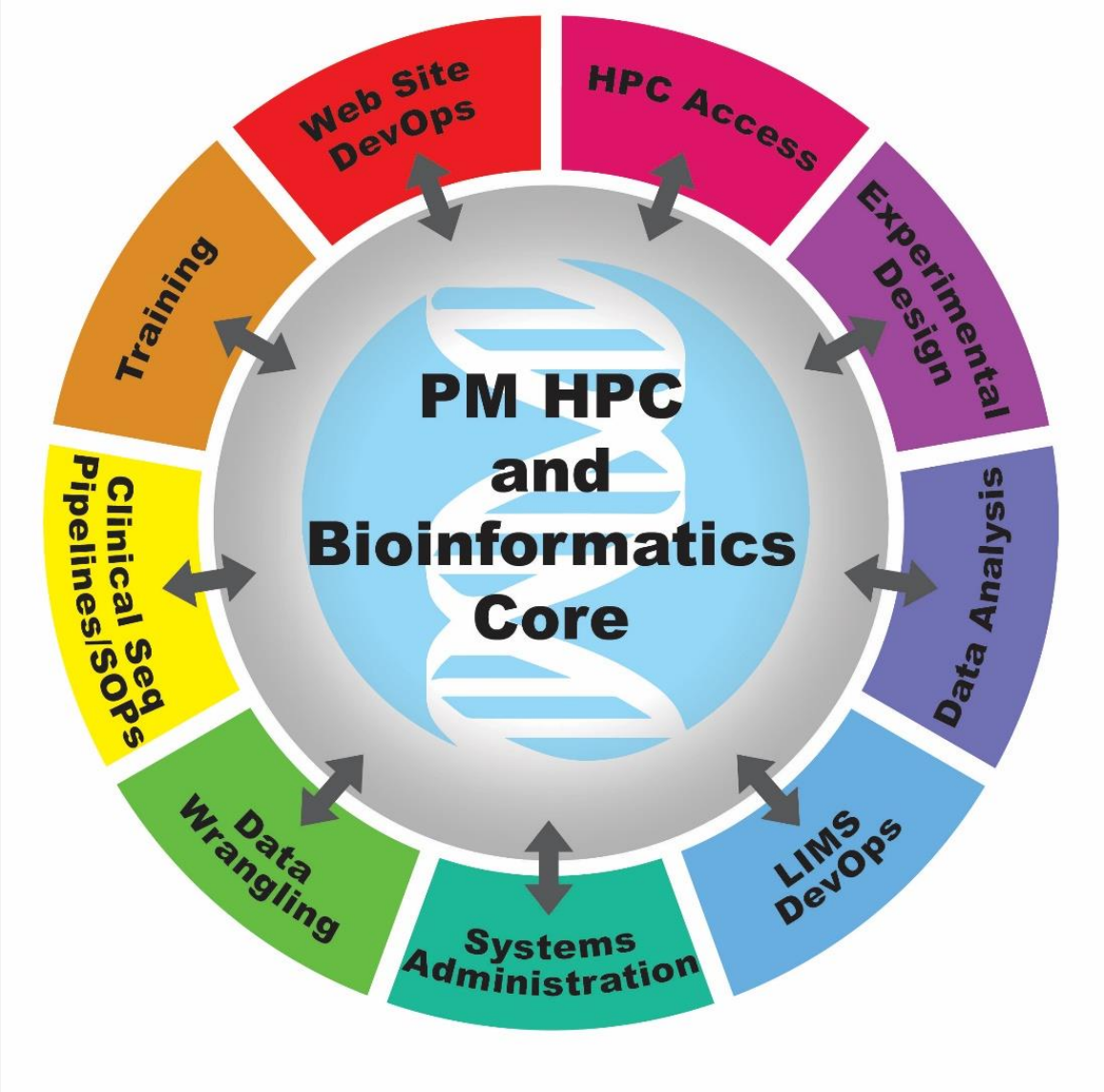

ACGT.../10001010...

THE DATA PROBLEM



Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4494865/

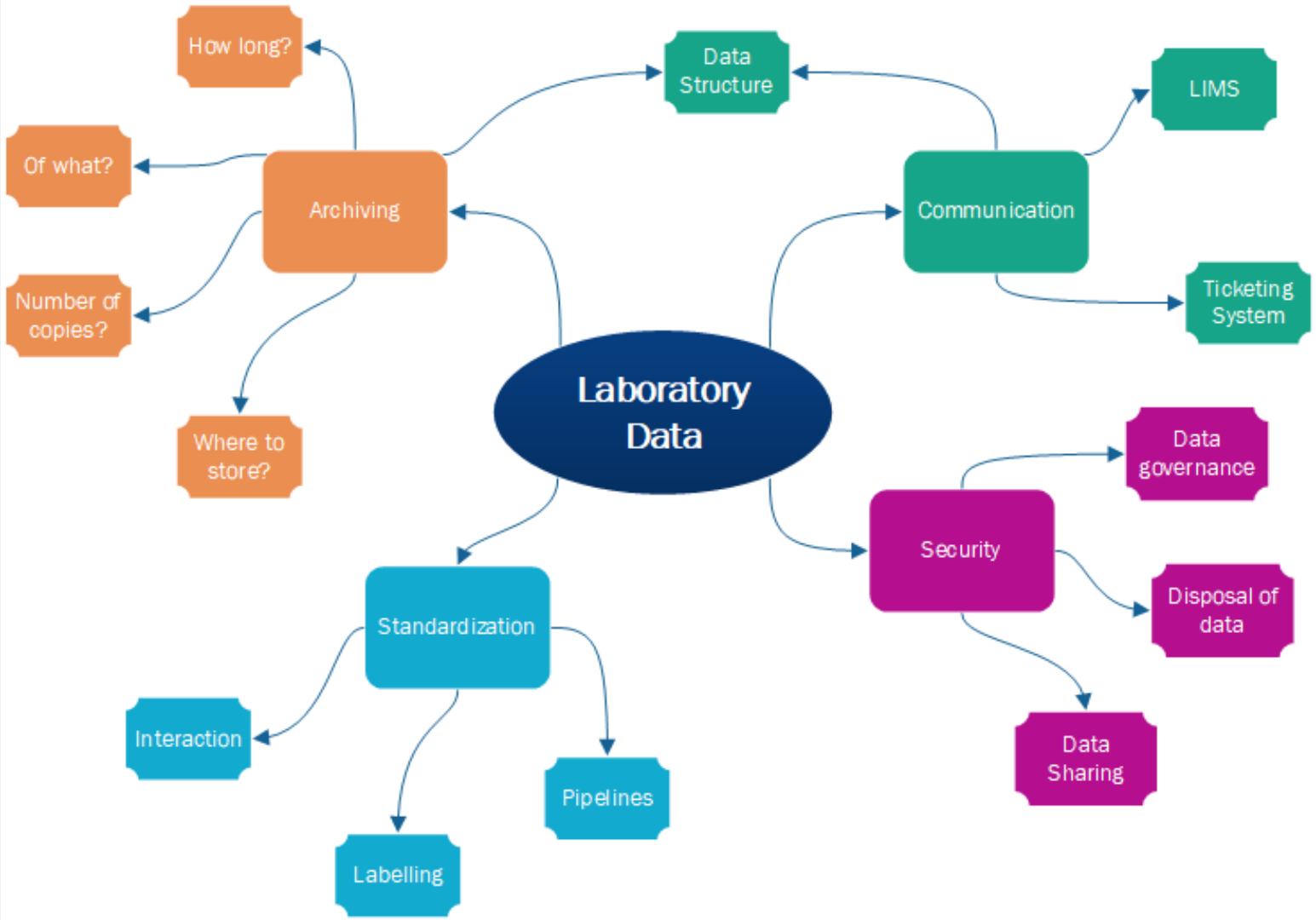
OUR CORE SERVICES



PURPOSE OF BIOINFORMATICS CORE

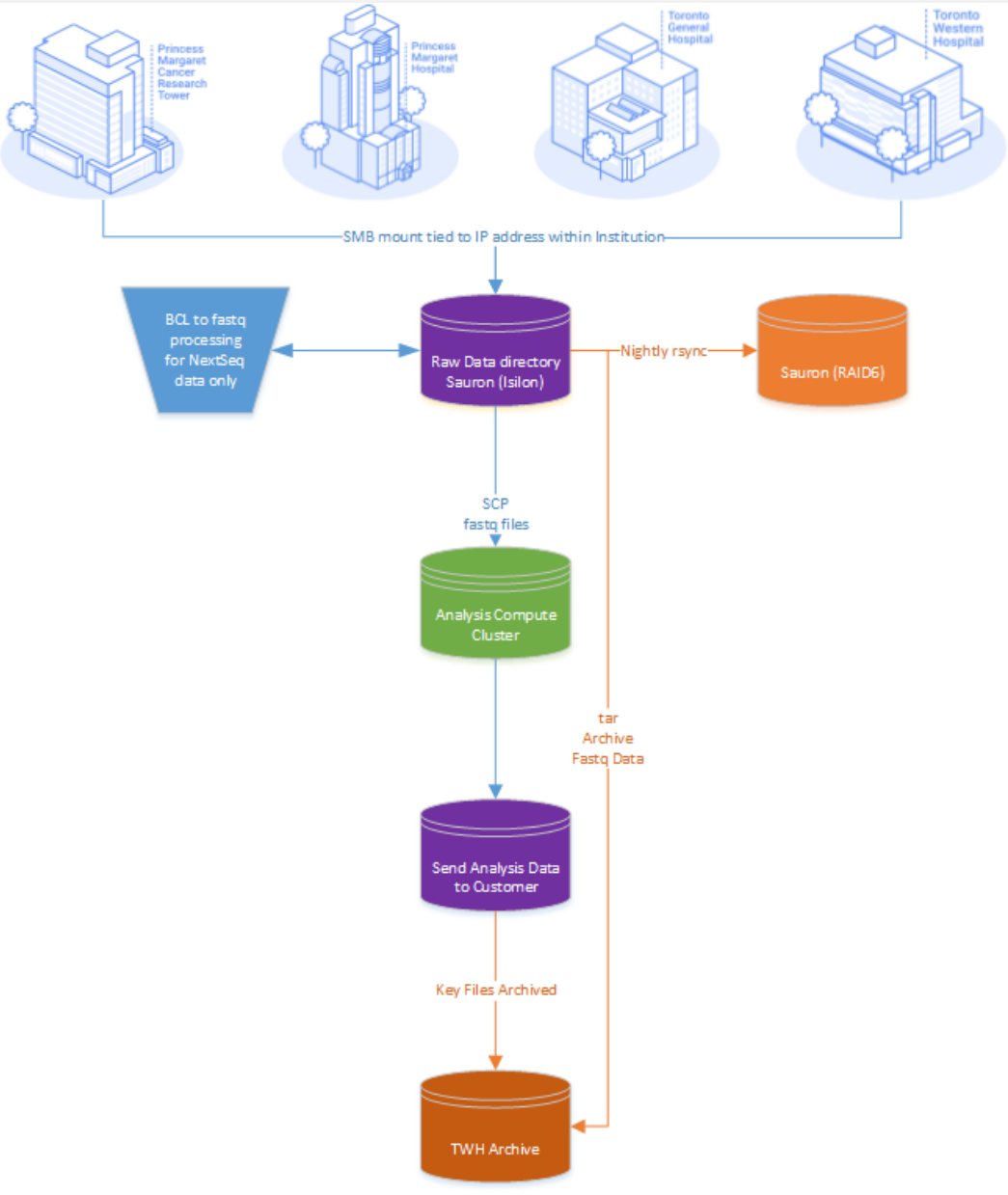
Factors to consider:

- Experimental Design
- Best practices for analysis from literature
- Need for HPC for large scale calculations
- Vast array of possible software options
- Implementation of governance for data (storage, security, access, sharing, etc.)



NGS DATA STORAGE & SECURITY

- All sequencers write directly to the core's storage
- Run is complete – data is moved to directory where initial processing can be completed – then sent to customer or further analyzed by the bioinformaticians in the core
- All raw data is backed up every day while in the initial storage directories
- Raw data is archived after initial run and analysis is completed
- If analysis is completed by the core's bioinformaticians key analysis files will also be archived.



INTERNAL HPC CAPABILITIES

- 1.1PB of high speed Infiniband distributed storage (Isilon)
- 200+TB of general purpose scratch (used for processing raw sequence data)
- ~1.4 PB of spinning disk backup at TWH (triple mirrored)
- 656 cores of CPU (384 Ivy Bridge/272 Sandy Bridge @ between 10 and 16GB of RAM/core)
- 10 Gigabit ethernet
- GPGPU node (nVidia Tesla M2090)
- Isolated web servers for web apps
- >120 users

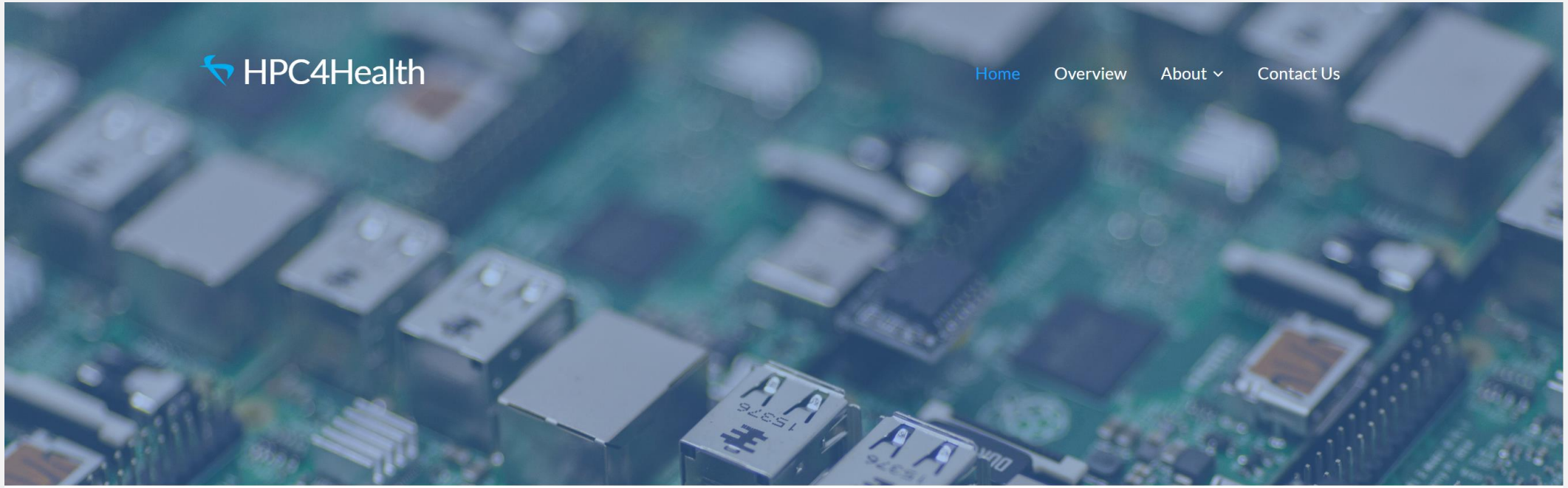
INTERNAL HPC CAPABILITIES CONT'D

- >200 TB of public data (some require approval eg. TCGA)
- Dozens of OS applications (Centos 6/7 series, module system)
- SGE, Slurm scheduler with multiple queues
- Custom scripts written for general use (pipelines for RNA-seq and DNA-seq)
- Command Line ONLY!

HPC4HEALTH

- ~2500 CPU cores latest Intel Xeon chips (>5000 total)
- 10 Gigabit ethernet
- 256 GB RAM/node
- 1.2 PB high-speed Isilon Storage
- ~20 Gb/s to data centre
- GPGPU nodes (4 nVidia Tesla P100 12GB each node)
- Torque/moab, Slurm scheduler

WHAT IS THE HPC4HEALTH?

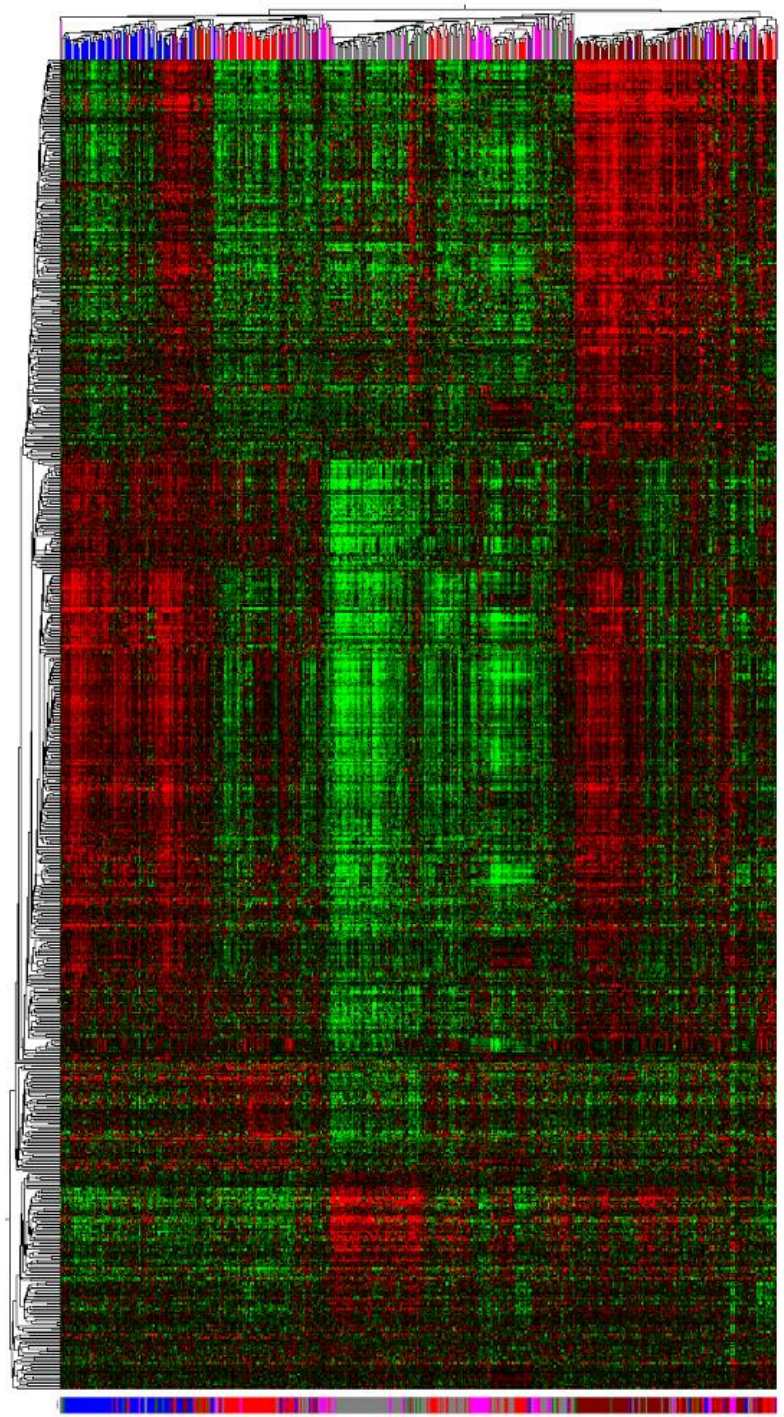


MISSION STATEMENT

Computing for Health

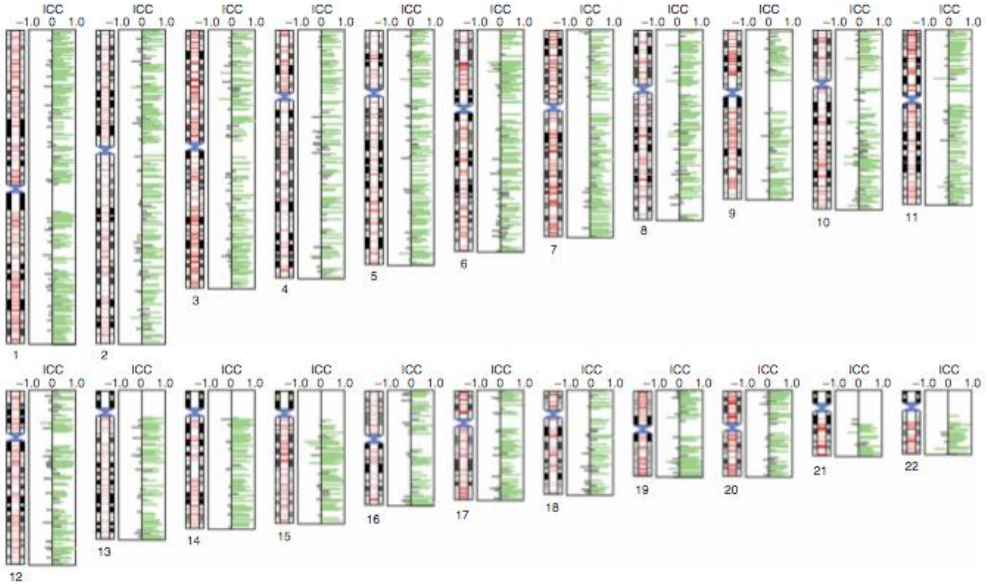
From genomics to medical imaging, almost every discipline in health-care is dealing with a "Data Deluge" of information. Translating this into something that will ultimately benefit patients requires massive amounts of computation and storage in an environment that is fast, secure, and run with efficiency. HPC4Health is a consortium of health providers who are working together to build the next-generation of compute engine for clinical research.

BIOINFORMATIC ANALYSES



Expression Signatures in 489 Ovarian Cancer Patients from TCGA

Genome-wide Methylation Patterns



ANNOTATED SEQUENCING RESULTS

Sequence Report Table for Patient74nf0.hardfiltered.noFail.avinput.hg19_multianno

Files included:Patient74nf0.hardfiltered.noFail.avinput.hg19_multianno.txt

Column Headings

Patient74nf0.hardfiltered.noFail.avinput.hg19_multianno.txt

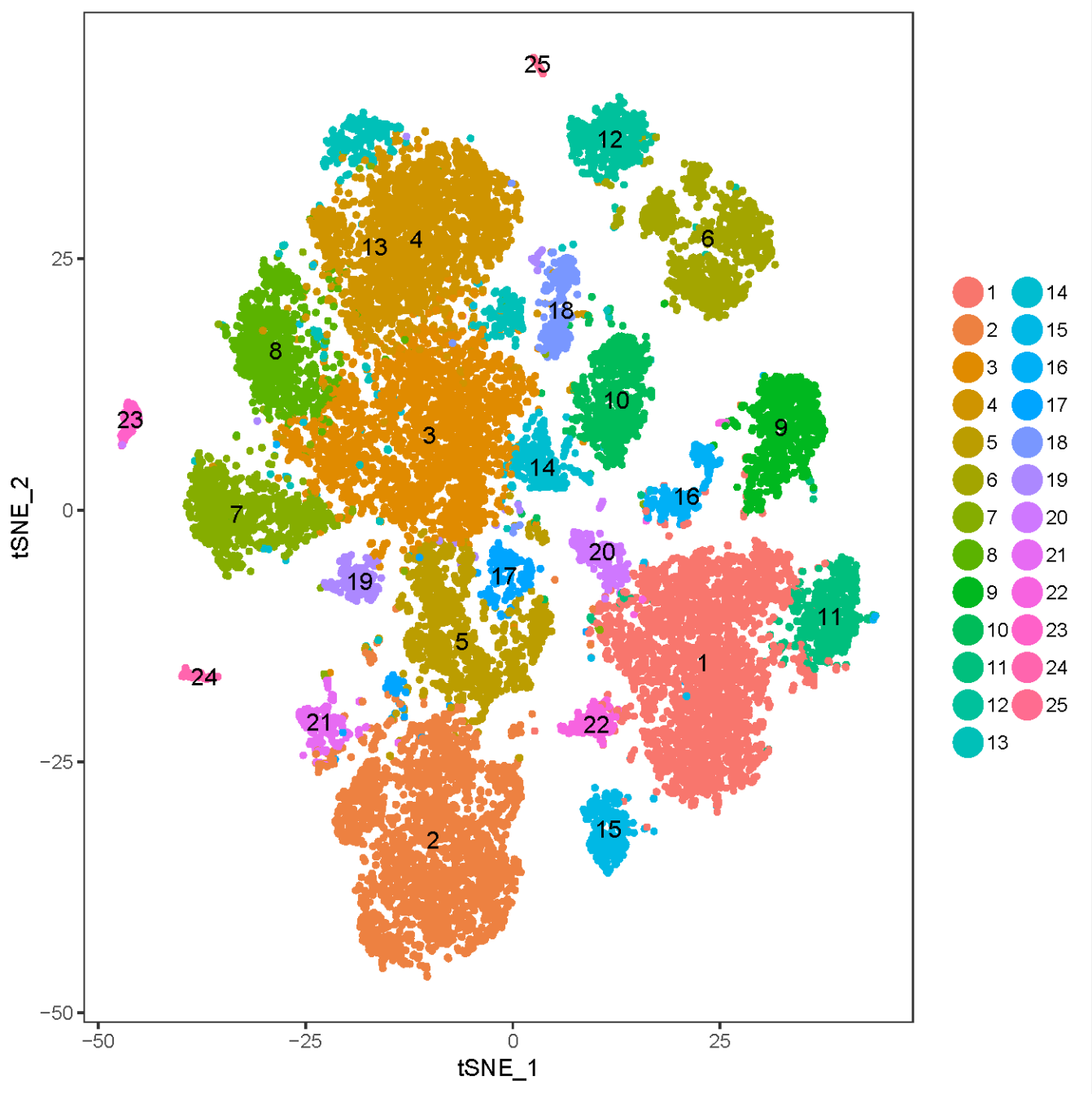
Chr	Start	End	Ref	Alt	Func.refGene	Gene.refG	ExonicFunc.refGene	AAChange.refGene	snp137	cosmic64	genomi	1000g2012apr_a	avsift	LJB_PhyloP	LJB_Pl	LJB_SIF	LJB_SI	LJB_PolyPhe	LJB_P(	LJB_GERP++	LJB_Mut	LJB_MutationTaster_Pre
chr1	205073034	205073034	G	C	exonic	RBBP5	nonsynonymous SNV	RBBP5:NM_001193272:exon5:c.C473G:p.S158C,RBBP5:NM_					0.13	0.999794	C	0.84	T	0.899	D	5.7	0.90917	D
chr1	114438951	114438951	A	G	exonic	AP4B1	nonsynonymous SNV	AP4B1:NM_001253852:exon8:c.T1439C:p.L480S,AP4B1:NM	rs1217401			0.35	0.13	0.999183	C	0.82	T	0.0010	B	5.69	0.73428	P
chr1	44877932	44877932	G	C	exonic	RNF220	nonsynonymous SNV	RNF220:NM_018150:exon2:c.G163C:p.D55H						0.999769	C	0.99	D	0.529	P	5.62	0.96010	D
chr1	103468308	103468308	T	G	exonic	COL11A1	nonsynonymous SNV	COL11A1:NM_080630:exon20:c.A1690C:p.N564H,COL11A1						0.998469	C	1.0	D	0.788621	NA	5.59	0.99853	D
chr1	113153597	113153597	T	C	exonic	ST7L	nonsynonymous SNV	ST7L:NM_017744:exon3:c.A317G:p.H106R,ST7L:NM_13872	rs143476452					0.998408	C	1.0	D	0.981	D	5.55	0.0	N
chr1	98348885	98348885	G	A	exonic	DPYD	nonsynonymous SNV	DPYD:NM_000110:exon2:c.C85T:p.R29C,DPYD:NM_001160	rs1801265			0.77	0.17	0.999726	C	0.98	D	0.713714	NA	5.49	0.71975	NA
chr1	84956143	84956143	C	T	exonic	RPF1	nonsynonymous SNV	RPF1:NM_025065:exon5:c.C604T:p.R202C	rs141904680			0.0014		0.999121	C	1.0	D	1.0	D	5.48	0.99999	D
chr1	46195375	46195375	T	C	exonic	IPP	nonsynonymous SNV	IPP:NM_001145349:exon4:c.A791G:p.K264R,IPP:NM_00589	rs28375469	ID=COSM		0.30	0.19	0.998274	C	0.59	T	0.0	B	5.46	0.97077	P
chr1	115168600	115168600	A	T	exonic	DENND2C	nonsynonymous SNV	DENND2C:NM_198459:exon2:c.T6A:p.D2E,DENND2C:NM_0	rs7541738			0.13	0.2	0.998841	C	0.11	T	0.07	B	5.36	0.12126	P
chr1	39908506	39908506	G	A	exonic	MACF1	nonsynonymous SNV	MACF1:NM_012090:exon75:c.G13048A:p.A4350T	rs587404			0.35	0.63	0.981078	C	0.42	T	0.235978	NA	5.35	8.0E-6	P
chr1	38265496	38265496	C	A	exonic	MANEAL	nonsynonymous SNV	MANEAL:NM_152496:exon2:c.C329A:p.S110Y,MANEAL:NM_						0.998899	C	1.0	D	0.895	D	5.27	0.99610	D
chr1	11150621	11150621	C	G	exonic	EXOSC10	nonsynonymous SNV	EXOSC10:NM_001001998:exon6:c.G748C:p.E250Q,EXOSC1					0.21	0.999348	C	0.74	T	0.024	B	5.23	0.93950	D
chr1	197087086	197087086	G	A	exonic	ASPM	stopgain SNV	ASPM:NM_001206846:exon17:c.C3898T:p.Q1300X,ASPM:NM						0.999584	C	0.90383	NA	0.735479	NA	5.18	1.0	A
chr1	62579891	62579891	G	T	exonic	INADL	nonsynonymous SNV	INADL:NM_176877:exon35:c.G4628T:p.G1543V	rs12140153			0.03		0.999582	C	1.0	D	0.853	D	5.15	0.99864	D
chr1	54606804	54606804	C	T	exonic	CDCP2	nonsynonymous SNV	CDCP2:NM_201546:exon3:c.G730A:p.G244R	rs3766465			0.83	0.07	0.99873	C	0.29	T	0.712	P	5.14	0.94867	P
chr1	114680310	114680310	G	A	exonic	SYT6	nonsynonymous SNV	SYT6:NM_001270805:exon3:c.C623T:p.T208I,SYT6:NM_205	rs41274114			0.17	0.38	0.999549	C	0.61	T	0.0030	B	5.1	0.99994	D
chr1	113252789	113252789	G	C	exonic	PPM1J	nonsynonymous SNV	PPM1J:NM_005167:exon10:c.C1514G:p.S505C						0.999529	C	1.0	D	0.6	P	5.07	0.28938	N
chr1	155255098	155255098	G	A	exonic	HCN3	nonsynonymous SNV	HCN3:NM_020897:exon5:c.G1232A:p.R411H	rs112470069			0.0005		0.999501	C	1.0	D	0.997	D	5.04	0.98423	D
chr1	25881403	25881403	G	A	exonic	LDLRAP1	nonsynonymous SNV	LDLRAP1:NM_015627:exon3:c.G284A:p.R95Q	rs139877665				0.12	0.999485	C	0.77	T	0.293	P	5.01	0.99977	D
chr1	27210721	27210721	T	C	exonic	GPN2	nonsynonymous SNV	GPN2:NM_018066:exon4:c.A790G:p.R264G	rs3170660			0.84	1	0.977717	C	0.0	T	0.0	B	5.01	6.0E-6	P
chr1	28282292	28282292	T	C	exonic	SMPDL3B	nonsynonymous SNV	SMPDL3B:NM_001009568:exon6:c.T788C:p.V263A,SMPDL3	rs3813803			0.22		0.997284	C	0.98	D	0.472	P	4.99	0.01425	P
chr1	89322997	89322997	G	A	exonic	GTF2B	nonsynonymous SNV	GTF2B:NM_001514:exon6:c.C709T:p.R237C	rs141868697			0.0009		0.999445	C	1.0	D	0.904	D	4.95	0.99999	D
chr1	118482149	118482149	C	T	exonic	WDR3	nonsynonymous SNV	WDR3:NM_006784:exon6:c.C629T:p.A210V	rs34973445			0.04	0.09	0.998462	C	0.95	T	0.0010	B	4.95	0.24837	N
chr1	94668200	94668200	C	A	exonic	ARHGAP29	nonsynonymous SNV	ARHGAP29:NM_004815:exon11:c.G1043T:p.R348L	rs140877322					0.987839	C	0.98	D	0.998	D	4.92	0.99877	D
chr1	72748140	72748140	G	C	exonic	NEGR1	nonsynonymous SNV	NEGR1:NM_173808:exon1:c.C38G:p.S13W						0.999408	C	0.98	D	0.998	D	4.9	0.90610	D
chr1	167097096	167097096	G	A	exonic	DUSP27	nonsynonymous SNV	DUSP27:NM_001080426:exon5:c.G2728A:p.D910N						0.99941	C	0.99	D	0.286	P	4.9	0.03992	N
chr1	12785494	12785494	G	T	exonic	AADACL3	nonsynonymous SNV	AADACL3:NM_001103169:exon3:c.G374T:p.C125F,AADACL	rs7513079			0.70	0.6	0.999395	C	0.18	T	0.280306	NA	4.89	0.77245	P
chr1	103379918	103379918	G	A	exonic	COL11A1	nonsynonymous SNV	COL11A1:NM_080630:exon50:c.C3620T:p.P1207L,COL11A1	rs3753841			0.53	0.12	0.999585	C	0.8	T	0.667388	NA	4.85	0.95499	P
chr1	24409191	24409191	C	T	exonic	MYOM3	nonsynonymous SNV	MYOM3:NM_152372:exon17:c.G1984A:p.G662R	rs4320729			0.32		0.998629	C	0.99	D	0.995	D	4.73	0.83218	P
chr1	36225948	36225948	T	C	exonic	CLSPN	nonsynonymous SNV	CLSPN:NM_001190481:exon8:c.A1574G:p.N525S,CLSPN:NM	rs7537203			0.46	0.05	0.996557	C	0.95	T	0.929	D	4.73	0.47967	P
chr1	23419856	23419856	G	A	exonic	LUZP1	nonsynonymous SNV	LUZP1:NM_001142546:exon3:c.C899T:p.T300I,LUZP1:NM_0					0.18	0.994788	C	0.91	T	0.96	D	4.69	0.09292	N
chr1	62732421	62732421	T	C	exonic	KANK4	nonsynonymous SNV	KANK4:NM_181712:exon6:c.A2302G:p.T768A	rs11207949			0.35	0.62	0.99729	C	0.78	T	0.497	P	4.66	0.53503	P
chr1	45224998	45224998	A	C	exonic	KIF2C	nonsynonymous SNV	KIF2C:NM_006845:exon14:c.A1345C:p.I449L	rs4342887			0.20		0.981636	C	0.99	D	0.043	B	4.64	0.99994	P

Search Refresh Select Columns Remove Rows Page 1 of 1 500 View 1 - 302 of 302

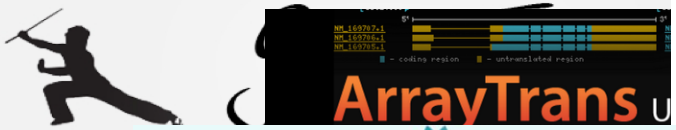
Generated by the OCI Genomics Centre on June 11 2013 at 15:40

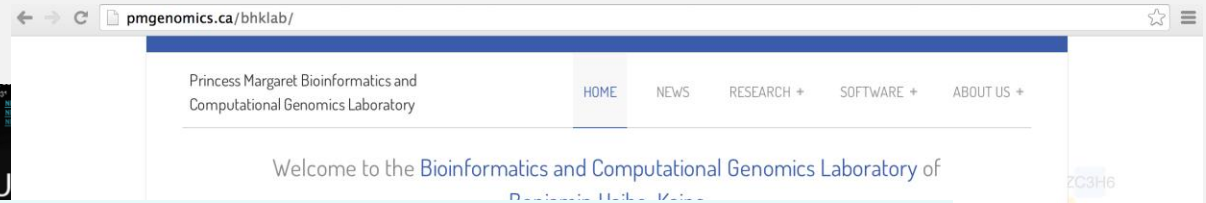
SINGLE CELL GENOMICS

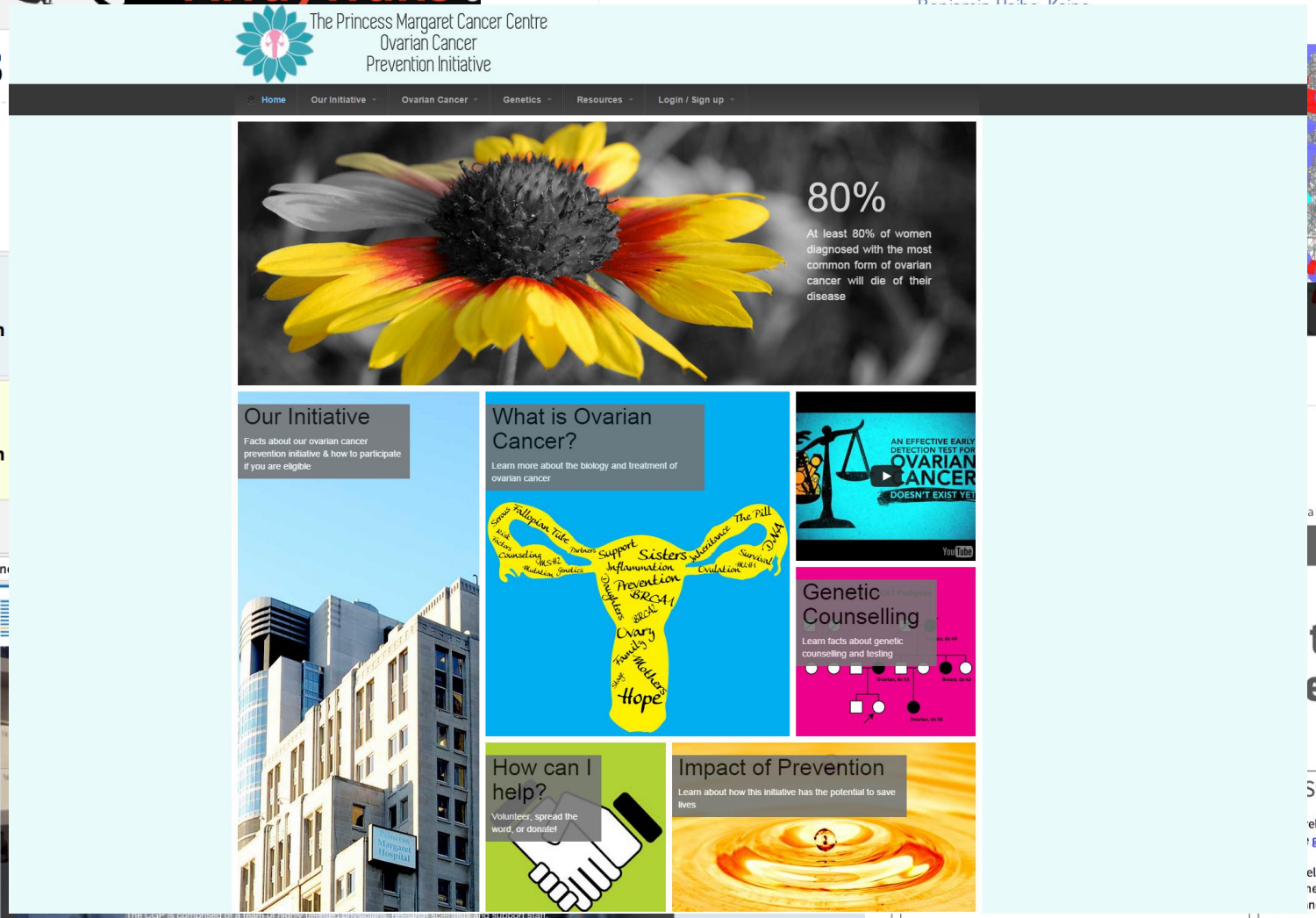
- Princess Margaret Genomics Centre is one of 2 in North America (only Canadian) 10x Genomics certified service provider
- Our core has extensive experience with this data

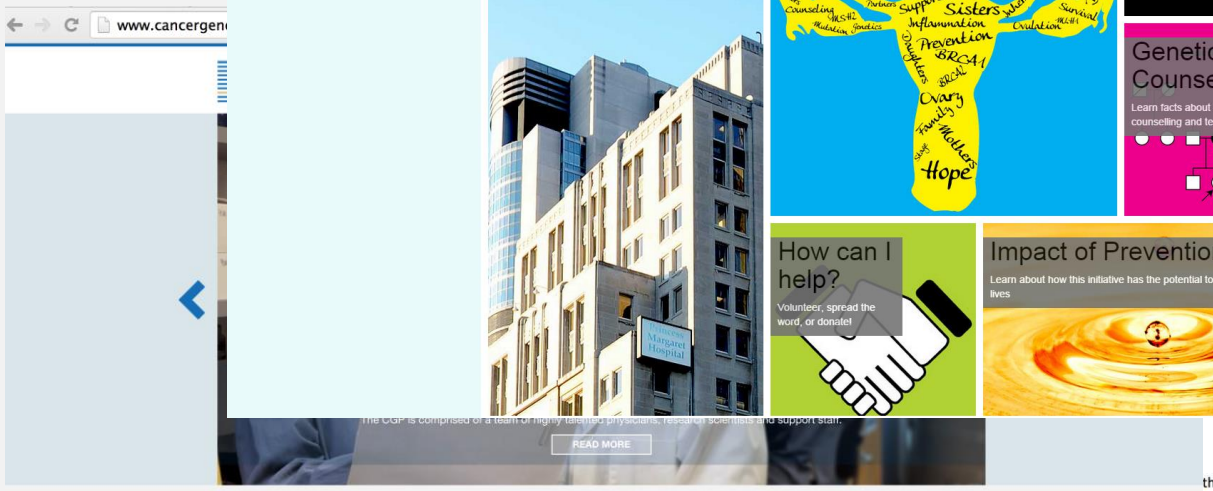


WEB SITES, WIKIS, CMS



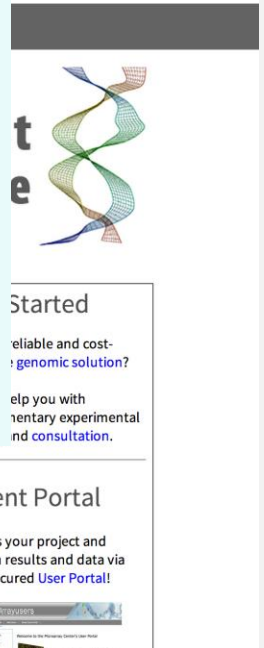






The **Cancer Genomics Program** at the Princess Margaret Cancer Centre is dedicated to advancing personalized cancer medicine through the identification of genetic mutations and molecular mechanisms that drive cancer.

Our ultimate vision is to have all patients who may benefit from this approach undergo comprehensive molecular characterization of their cancer by 2018.



WIKIS

Carl Talk Preferences Watchlist Contributions Log out

PRINCESS MARGARET GENOMICS CENTRE

Pmgc Discussion Read Edit View history Search

- Home
  - Main
  - Event Calendar
- Data Analysis
  - High Throughput Sequencing
  - Microarray
  - Software
- System
- Web
  - Web Sites
  - Web Servers
  - Software
  - Document
- Database
- Repository
- Backup
- Help
- Tools

Pmgc:HTS RNASeq

Contents [hide]

Tuxedo Package [edit]

Workflow using the tuxedo package can be found at Cufflinks website.

- scripts and wrappers are located under /mnt/work1/software/ocigcscripts/production/tophat2_wrapper
- There is a "qsub" sub-directory under Output directory, and they are real scripts to perform the jobs. You can repeat the job by "qsub" each script.
- By default, there is a log file under output directory called 'process_log.txt'.

Raw Data Structure [edit]

- Fastq files for each sample should be in a directory named "Sample_<sample name>"
- Fastq files from multiple lanes for the same sample should be put into the same directory.
- Fastq files can be gzipped or unzipped.
- Fastq file name format "<sample name>_<barcode>_<lane number>"

Example:

```
fastq
├── Sample_D001-Blood
|   ├── D001-Blood_AGAGTC_L001_R1.fastq
```

TRACK EVERYTHING!

141205_SN1080_0189_BC6345ACXX [edit]

- 51 cycles + 7 indexing cycles
- sample sheet
- bcl2fastq with mismatch=1, barcode=6
- barcode error for sample 3088 on lane 6. Changed to CAGATC and re-ran the conversion.
- Project_DeCarvalho_Hung (Roxana)
  - lane 1-4
  - fastq files were copied into /mnt/work1/users/decarvalhogroup/141205_SN1080_0189_BC6345ACXX_Hung on Dec 10
- Project_DeCarvalho (Tiago)
  - lane 5-8
  - fastq files were copied into /mnt/work1/users/lupiengroup/141205_SN1080_0189_BC6345ACXX on Dec 10

141204_SN1080_0188_AC5V7DACXX [edit]

- 51 cycles + 8 indexing cycles
- sample sheet
- bcl2fastq with mismatch=1, barcode=6
- Project_Dick
  - lane 1-6,8
  - fastq files were copied into /mnt/work1/users/lupiengroup/141204_SN1080_0188_AC5V7DACXX_Naoya. Emailed Naoya tnaoya19760517@gmail.com and Nadia npenrod@uhnresearch.ca on Dec 8.
- Project_DeCarvlho_Hung
  - lane 7
  - fastq files were copied into /mnt/work1/users/decarvalhogroup/141204_SN1080_0188_AC5V7DACXX_Hung

141120_SN1068_0163_AC5FTVACXX [edit]

- 51 cycles + 9 indexing cycles
- sample sheet
- bcl2fastq with mismatch=1, barcode=6
- Project_Lupien
  - lane 1-4
  - fastq files were copied into /mnt/work1/users/lupiengroup/141120_SN1068_0163_AC5FTVACXX_ken on Nov 25.
- Project_Dick
  - lane 5-8
  - fastq files were copied into /mnt/work1/users/lupiengroup/141120_SN1068_0163_AC5FTVACXX_Dick on Nov 25.

ALIGNMENT - CANADIAN BIOINFORMATICS

- Canadian Centre for Computational Genomics (C3G)
- Canadian Bioinformatics Workshops (CBW)
- Compute Canada Bioinformatics Helpdesk
- CanDIG (http://CanDIG.github.io)
- AACR Project GENIE

BIOINFORMATICS & HPC CORE

Director
Carl Virtanen[RC]

Manager
Natalie Stickle[RC]

System Administration Manager
Zhibin Lu[RC]
Analyst
Qun Jin[RC]

Bioinformaticians

Clinical
Irene Chae[C]
Roozbeh Dolatshahi[C]
Gregory Downs[C]

Research
Richard De Borja[RC]
Abdellatif Daghrach[RC]
Neke Ibeh[R]