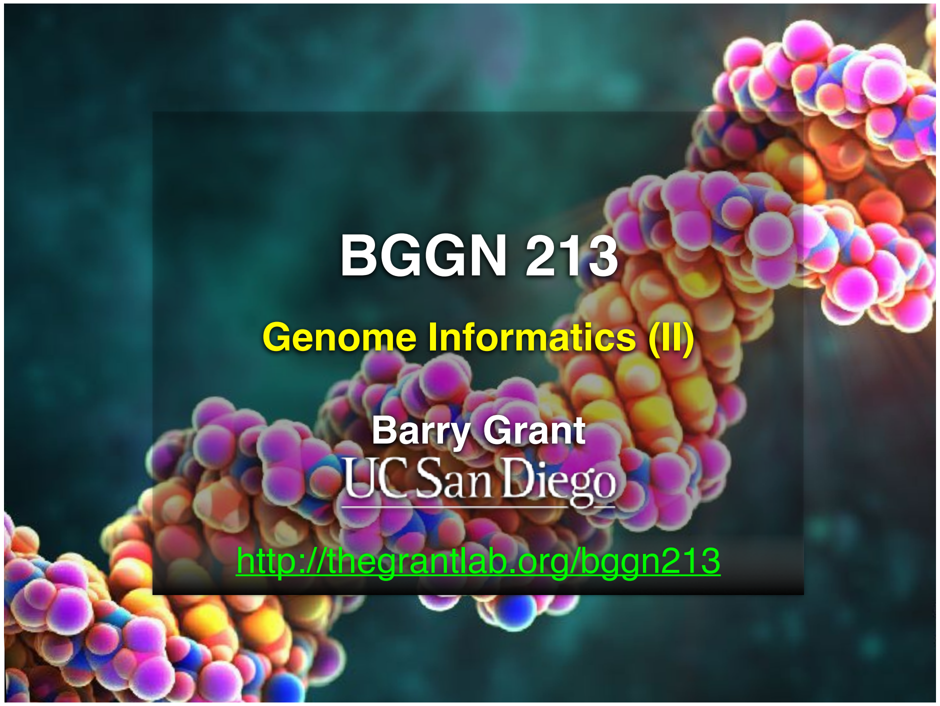


BGGN 213

Genome Informatics (II)

Barry Grant

UC San Diego

http://thegrantlab.org/bggn213

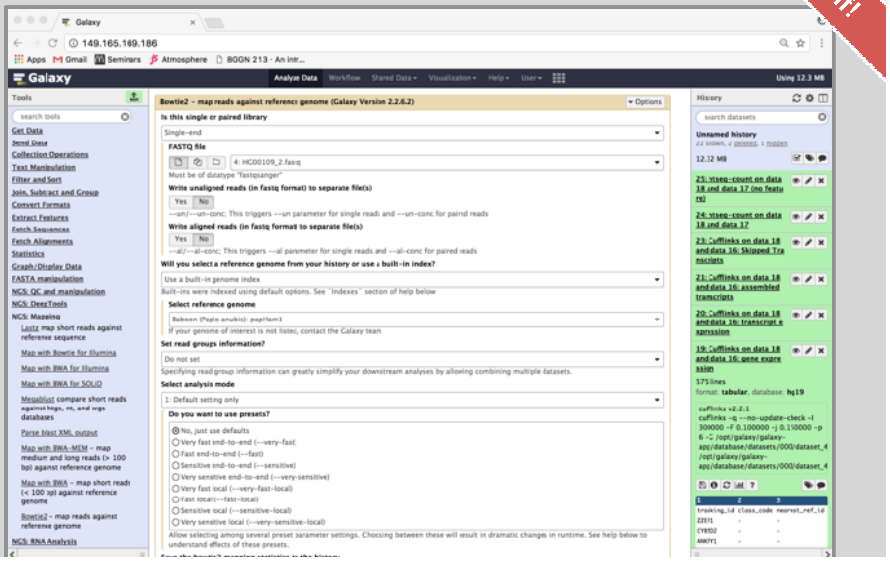
TODAYS MENU:

- **What is a Genome?**
  - Genome sequencing and the Human genome project
- **What can we do with a Genome?**
  - Comparative genomics
- **Modern Genome Sequencing**
  - 1st, 2nd and 3rd generation sequencing
- **Workflow for NGS**
  - RNA-Sequencing and discovering variation

Start a jetstream galaxy instance!

Do it Yourself!

http://tinyurl.com/bggn213-L15


Galaxy
149.165.169.186
Apps
Gmail
Seminars
Atmosphere
BGGN 213 - An intr...
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 12.3 MB
Tools
search tools
Get Data
Send Data
Collection Operations
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data
FASTA manipulation
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
Lastz map short reads against reference sequence
Map with Bowtie for Illumina
Map with BWA for Illumina
Map with BWA for SOLiD
Megablast compare short reads against htgs, nt, and wgs databases
Parse blast XML output
Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome
Map with BWA - map short reads (< 100 bp) against reference genome
Bowtie2 - map reads against reference genome
NGS: RNA Analysis
Options
History
search datasets
Unnamed history
12.12 MB
25: htseq-count on data 18 and data 17 (no feature)
24: htseq-count on data 18 and data 17
23: Cufflinks on data 18 and data 16: Skipped Transcripts
21: Cufflinks on data 18 and data 16: assembled transcripts
20: Cufflinks on data 18 and data 16: transcript expression
19: Cufflinks on data 18 and data 16: gene expression
575 lines
format: tabular, database: hg19

Bowtie2 - map reads against reference genome (Galaxy Version 2.2.6.2)

Is this single or paired library: Single-end

FASTQ file: 4: HG00109_2.fastq

Must be of datatype "fastqsanger"

Write unaligned reads (in fastq format) to separate file(s): Yes [✓ No]

--un/--un-conc; This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s): Yes [✓ No]

--al/--al-conc; This triggers --al parameter for single reads and --al-conc for paired reads

Will you select a reference genome from your history or use a built-in index?: Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Select reference genome: Baboon (Papio anubis): papHam1

If your genome of interest is not listed, contact the Galaxy team

Set read groups information?: Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode: 1: Default setting only

Do you want to use presets?

● No, just use defaults
○ Very fast end-to-end (--very-fast)
○ Fast end-to-end (--fast)
○ Sensitive end-to-end (--sensitive)
○ Very sensitive end-to-end (--very-sensitive)
○ Very fast local (--very-fast-local)
○ Fast local (--fast-local)
○ Sensitive local (--sensitive-local)
○ Very sensitive local (--very-sensitive-local)

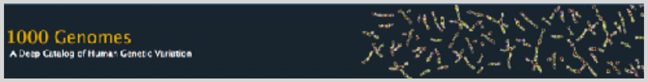
Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets.

```
cufflinks v2.2.1
cufflinks -q --no-update-check -I 300000 -F 0.100000 -j 0.150000 -p 6 -G /opt/galaxy/galaxy-app/database/datasets/000/dataset_4
/opt/galaxy/galaxy-app/database/datasets/000/dataset_4
```

1	2	3
tracking_id	class_code	nearest_ref_id
ZZEF1	-	-
CYB5D2	-	-
ANKFY1	-	-

Population Scale Analysis

We can now begin to assess genetic differences on a very large scale, both as naturally occurring variation in human and non-human populations as well somatically within tumors



The Cancer Genome Atlas
Understanding genomics to improve cancer care

“Variety’s the very spice of life”

–William Cowper, 1785

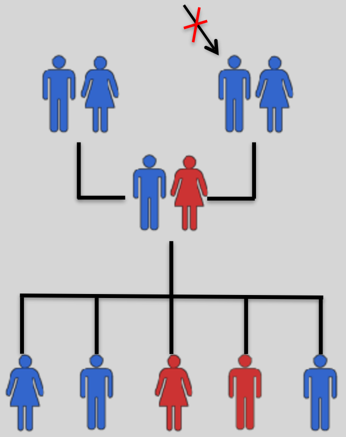
“Variation is the spice of life”

–Kruglyak & Nickerson, 2001

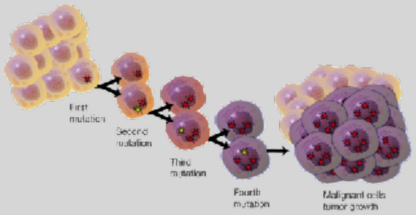
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility to disease, or be completely benign.

Germline Variation

- Mutations in the germline are passed along to offspring and are present in the DNA over every cell
- In animals, these typically occur in meiosis during gamete differentiation



Somatic Variation



- Mutations in non-germline cells that are not passed along to offspring
- Can occur during mitosis or from the environment itself
- Are an integral part in tumor progression and evolution

Darryl Leja, Courtesy: National Human Genome Research Institute.

Mutation vs Polymorphism

- A mutation must persist to some extent within a population to be considered polymorphic
  - >1% frequency is often used
- Germline mutations that are not polymorphic are considered rare variants

> *“From the standpoint of the neutral theory, the rare variant alleles are simply those alleles whose frequencies within a species happen to be in a low-frequency range (0,q), whereas polymorphic alleles are those whose frequencies happen to be in the higher-frequency range (q, 1-q), where I arbitrarily take q = 0.01. Both represent a phase of molecular evolution.”*
>
> *-Motoo Kimura*

Kimura M (1983) Mol. Biol. Evol., 1(1), pp. 84-93

Types of Genomic Variation

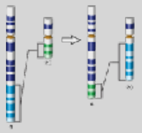
- **Single Nucleotide Polymorphisms** (SNPs) – mutations of one nucleotide to another

```
AATCTGAGGCAT
AATCTCAGGCAT
```

- **Insertion/Deletion Polymorphisms** (INDELs) – small mutations removing or adding one or more nucleotides at a particular locus

```
AATCTGAAGGCAT
AATCT--AGGCAT
```

- **Structural Variation** (SVs) – medium to large sized rearrangements of chromosomal DNA



Darryl Leja, Courtesy: National Human Genome Research Institute.

Differences Between Individuals

The average number of genetic differences in the germline between two random humans can be broken down as follows:

- 3,600,000 single nucleotide differences
- 344,000 small insertion and deletions
- 1,000 larger deletion and duplications

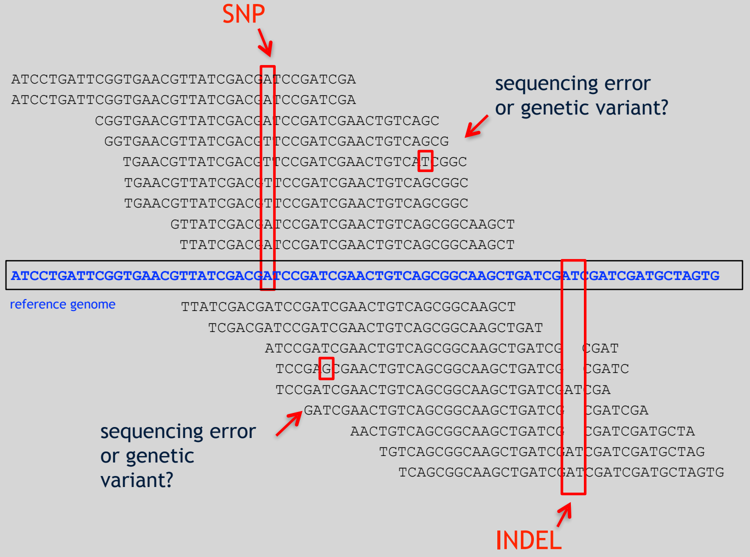
Numbers change depending on ancestry!

1000 Genomes Project, Nature, 2012

Discovering Variation: SNPs and INDELs

- Small variants require the use of sequence data to initially be discovered
- Most approaches align sequences to a reference genome to identify differing positions
- The amount of DNA sequenced is proportional to the number of times a region is covered by a sequence read
  - More sequence coverage equates to more support for a candidate variant site

Discovering Variation: SNPs and INDELs

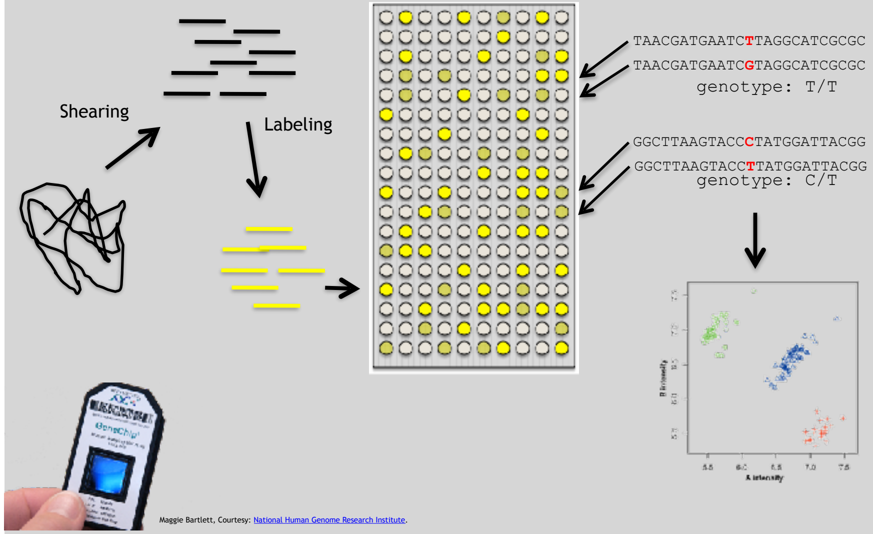

SNP
sequencing error or genetic variant?
reference genome
sequencing error or genetic variant?
INDEL

```
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
         CGGTGAACGTTATCGACGATCCGATCGAACTGTCAGC
          GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
            TGAACGTTATCGACGTTCCGATCGAACTGTCATCGGC
            TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
            TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
                 GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
                  TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG
                  TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
                     TCGACGATCCGATCGAACTGTCAGCGGCAAGCTGAT
                           ATCCGATCGAACTGTCAGCGGCAAGCTGATCG  CGAT
                            TCCGAGCGAACTGTCAGCGGCAAGCTGATCG  CGATC
                            TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGA
                               GATCGAACTGTCAGCGGCAAGCTGATCG  CGATCGA
                                    AACTGTCAGCGGCAAGCTGATCG  CGATCGATGCTA
                                       TGTCAGCGGCAAGCTGATCGATCGATCGATGCTAG
                                         TCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG
```

Genotyping Small Variants

- Once discovered, oligonucleotide probes can be generated with each individual allele of a variant of interest
- A large number can then be assessed simultaneously on microarrays to detect which combination of alleles is present in a sample

SNP Microarrays



Shearing

Labeling

TAACGATGAATC**T**TAGGCATCGCGC

TAACGATGAATC**G**TAGGCATCGCGC

genotype: T/T

GGCTTAAGTACC**C**TATGGATTACGG

GGCTTAAGTACC**T**TATGGATTACGG

genotype: C/T

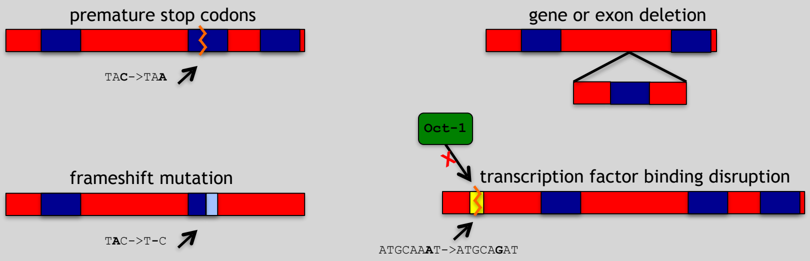
Maggie Bartlett, Courtesy: National Human Genome Research Institute.

Discovering Variation: SVs

- Structural variants can be discovered by both sequence and microarray approaches
- Microarrays can only detect genomic imbalances, specifically copy number variants (CNVs)
- Sequence based approaches can, in principle, identify all types of structural rearrangements

Impact of Genetic Variation

There are numerous ways genetic variation can exhibit functional effects



premature stop codons

TA**C**->TA**A**

gene or exon deletion

frameshift mutation

T**A**C->T-C

Oct-1

transcription factor binding disruption

ATGCAA**A**T->ATGCA**G**AT

Variant Annotation

- Variants are *annotated* based on their potential functional impact
- For variants falling inside genes, there are a number of software packages that can be used to quickly determine which may have a functional role (missense/ nonsense mutations, splice site disruption, etc)
- A few examples are:
  - ANNOVAR (http://www.openbioinformatics.org/annovar/)
  - VAAST (http://www.yandell-lab.org/software/vaast.html)
  - VEP (http://http://grch37.ensembl.org/Homo_sapiens/Tools/VEP)
  - SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation134/)
  - snpEff (http://snpeff.sourceforge.net/)

Variant Annotation Classes

High Impact

- exon_deleted
- frame_shift
- splice_acceptor
- splice_donor
- start_loss
- stop_gain
- stop_loss
- non_synonymous_start
- transcript_codon_change

Medium Impact

- non_syn_coding
- inframe_codon_gain
- inframe_codon_loss
- inframe_codon_change
- codon_change_del
- codon_change_ins
- UTR_5_del
- UTR_3_del
- other_splice_variant
- mature_miRNA
- regulatory_region
- TF_binding_site
- regulatory_region_ablation
- regulatory_region_amplification
- TFBS_ablation
- TFBS_amplification

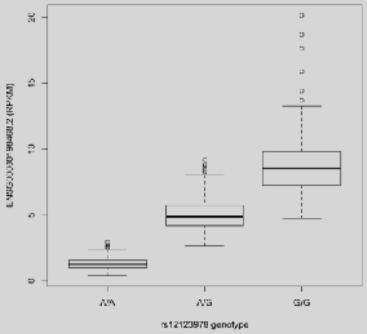
Low Impact

- synonymous_stop
- synonymous_coding
- UTR_5_prime
- UTR_3_prime
- intron
- CDS
- upstream
- downstream
- intergenic
- intragenic
- gene
- transcript
- exon
- start_gain
- synonymous_start
- intron_conserved
- nc_transcript
- NMD_transcript
- transcript_codon_change
- incomplete_terminal_codon
- nc_exon
- transcript_ablation
- transcript_amplification
- feature elongation
- feature truncation

GEMINI, http://gemini.readthedocs.org/

Variation and Gene Expression

- Expression quantitative trait loci (eQTLs) are regions of the genome that are associated with expression levels of genes
- These regions can be nearby (cis) or far away (trans) from the genes that they affect
- Genetic variants in eQTL regions are typically responsible through changes to regulatory elements



Data generated from http://www.geuvadis.org/

Geuvadis Consortium

http://www.geuvadis.org/web/geuvadis

GEUVADIS Genetic European Variation in Health and Disease, A European Medical Sequencing Consortium

GEUVADIS RNA sequencing project for 1000 Genomes samples

Welcome!

Welcome to the GEUVADIS website

Funded by the European Commission (FP7), GEUVADIS brings together [illegible]

Upcoming Geuvadis Events

Genomic Medicine in the Mediterranean

Inaugural conference

GENOMIC MEDICINE IN THE MEDITERRANEAN (GM²) INAUGURAL CONFERENCE OCTOBER 2-5, 2013

Additional Reference Slides
on FASTQ format, ASCII Encoded Base Qualities,
FastQC, Alignment and SAM/BAM formats

More fu

Raw data usually in **FASTQ format**

```
@NS500177:196:HFTTTAFXX:1:11101:10916:1458 2:N:0:CGCGGCTG                 1
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAAGCAGCCGGTGTAA  2
+                                                                         3
AAAAAEEEEEEEEEEEE//AEEEAEEEEEEEEEEEE/EE/<<EE/AAEEAEE///EEEEAEEEAEA<   4
```

Each sequencing "read" consists of 4 lines of data :

1 The first line (which always starts with '**@**') is a unique ID for the sequence that follows

2 The second line contains the bases called for the sequenced fragment

3 The third line is always a "+" character

4 The forth line contains the quality scores for each base in the sequenced fragment (these are ASCII encoded...)

ASCII Encoded Base Qualities

```
@NS500177:196:HFTTTAFXX:1:11101:10916:1458 2:N:0:CGCGGCTG
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAAGCAGCCGGTGTAA
+
AAAAAEEEEEEEEEEEE//AEEEAEEEEEEEEEEEE/EE/<<EE/AAEEAEE///EEEEAEEEAEA<   4
```

- Each sequence base has a corresponding numeric quality score encoded by a single ASCII character typically on the 4th line (see 4 above)
- ASCII characters represent integers between 0 and 127
- Printable ASCII characters range from 33 to 126
- Unfortunately there are 3 quality score formats that you may come across...

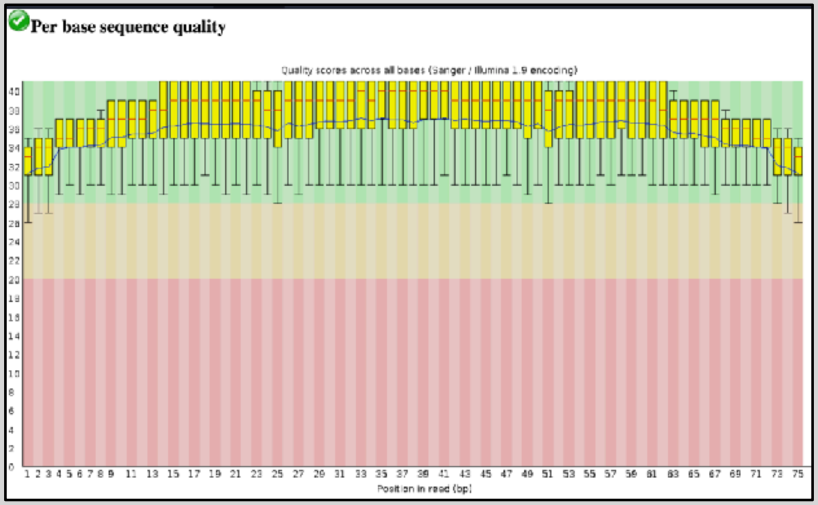
Interpreting Base Qualities in R

	ASCII Range	Offset	Score Range
Sanger, Illumina (Ver > 1.8)	**33-126**	**33**	**0-93**
Solexa, Ilumina (Ver < 1.3)	59-126	64	5-62
Illumina (Ver 1.3 -1.7)	64-126	64	0-62

```
> library(seqinr)
> library(gtools)
> phred <- asc( s2c("DDDDCDEDCDDDDBBDDDCC@") ) - 33
> phred
##  D  D  D  D  C  D  E  D  C  D  D  D  D  B  B  D  D  D  C  C  @
## 35 35 35 35 34 35 36 35 34 35 35 35 35 33 33 35 35 35 34 34 31

> prob <- 10**(-phred/10)
```

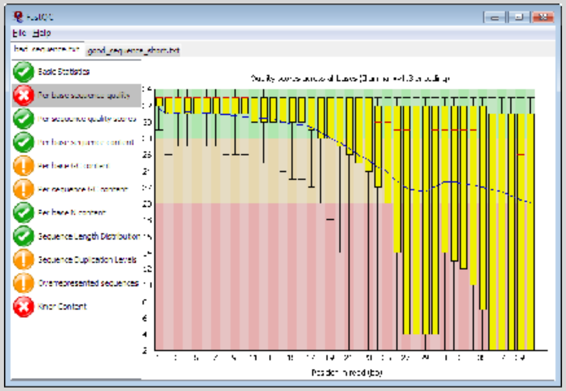
FastQC Report



FASTQC

FASTQC is one approach which provides a visual interpretation of the raw sequence reads

- http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



Sequence Alignment

- Once sequence quality has been assessed, the next step is to align the sequence to a reference genome
- There are *many* distinct tools for doing this; which one you choose is often a reflection of your specific experiment and personal preference

BWA
Bowtie
SOAP2
Novoalign
mr/mrsFast
Eland
Blat
Bfast

BarraCUDA
CASHx
GSNAP
Mosiak
Stampy
SHRiMP
SeqMap
SLIDER

RMAP
SSAHA
etc

SAM Format

- Sequence Alignment/Map (**SAM**) format is the almost-universal sequence alignment format for NGS
  - binary version is BAM
- It consists of a header section (lines start with '@') and an alignment section
- The official specification can be found here:
  - http://samtools.sourceforge.net/SAM1.pdf

Example SAM File

Header section

```
@HD                 VN:1.0              SO:coordinate
@SQ                 SN:1                LN:249250621        AS:NCBI37           UR:file:/data/local/ref/GATK/human_g1k_v37.fasta            M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ                 SN:2                LN:243199373        AS:NCBI37           UR:file:/data/local/ref/GATK/human_g1k_v37.fasta            M5:a0d9851da00400dec1098a9255ac712e
@SQ                 SN:3                LN:198022430        AS:NCBI37           UR:file:/data/local/ref/GATK/human_g1k_v37.fasta            M5:fdfd811849cc2fadebc929bb925902e5
@RG                 ID:UM0098:1         PL:ILLUMINA         PU:HWUSI-EAS1707-615LHAAXX-L001         LB:80               DT:2010-05-05T20:00:00-0400             SM:SD37743          CN:UMCORE
@RG                 ID:UM0098:2         PL:ILLUMINA         PU:HWUSI-EAS1707-615LHAAXX-L002         LB:80               DT:2010-05-05T20:00:00-0400             SM:SD37743          CN:UMCORE
@PG                 ID:bwa              VN:0.5.4
```

Alignment section

```
1:497:R:-272+13M17D24M                  113                 1                   497                 37                  37M                 15                  100338662           0
                    CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG   0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>   XT:A:U              NM:i:0              SM:i:37             AM:i:0              X0:i:1
                    X1:i:0              XM:i:0              XO:i:0              XG:i:0              MD:Z:37
19:20389:F:275+18M2D19M                 99                  1                   17644               0                   37M                 =                   17919               314
                    TATGACTGCTAATAATACCTACACATGTTAGAACCAT   >>>>>>>>>>>>>>>>>>>><<>>><<>>4::>>:<9   RG:Z:UM0098:1       XT:A:R              NM:i:0              SM:i:0              AM:i:0
                    X0:i:4              X1:i:0              XM:i:0              XO:i:0              XG:i:0              MD:Z:37
19:20389:F:275+18M2D19M                 147                 1                   17919               0                   18M2D19M            =                   17644               -314
                    GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT   ;44999;499<8<8<<<8<<><<<<><7<;<<<><<    XT:A:R              NM:i:2              SM:i:0              AM:i:0              X0:i:4
                    X1:i:0              XM:i:0              XO:i:1              XG:i:2              MD:Z:18^CA19
9:21597+10M2I25M:R:-209                 83                  1                   21678               0                   8M2I27M             =                   21469               -244
                    CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT   <;9<<5><<<<><<<><<><>><9>><>>>9>>><>    XT:A:R              NM:i:2              SM:i:0              AM:i:0              X0:i:5
                    X1:i:0              XM:i:0              XO:i:1              XG:i:2              MD:Z:35
```

http://genome.sph.umich.edu/wiki/SAM

SAM Utilities

- **Samtools** is a common toolkit for analyzing and manipulating files in SAM/ BAM format
  - http://samtools.sourceforge.net/
- **Picard** is a another set of utilities that can used to manipulate and modify SAM files
  - http://picard.sourceforge.net/
- These can be used for viewing, parsing, sorting, and filtering SAM files as well as adding new information (e.g. Read Groups)

Genome Analysis Toolkit (**GATK**)

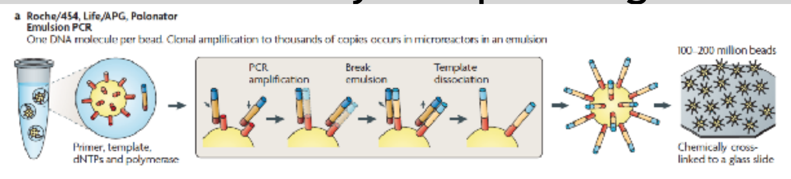
- Developed in part to aid in the analysis of 1000 Genomes Project data
- Includes many tools for manipulating, filtering, and utilizing next generation sequence data
- http://www.broadinstitute.org/gatk/

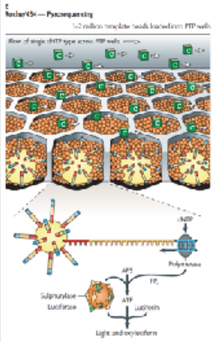
Do it Yourself!

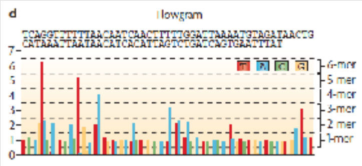
Additional Reference Slides on Sequencing Methods

Hands-on worksheet:
http://tinyurl.com/bggn213-L15

Roche 454 - Pyrosequencing

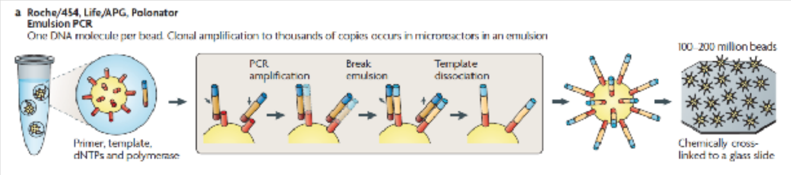

a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Primer, template, dNTPs and polymerase
Chemically cross-linked to a glass slide

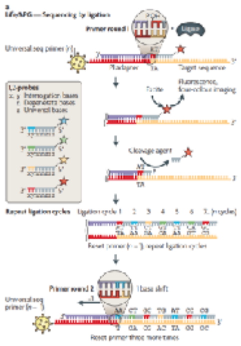


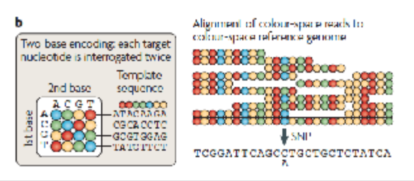


Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46

Life Technologies SOLiD – Sequence by Ligation


a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Primer, template, dNTPs and polymerase
Chemically cross-linked to a glass slide

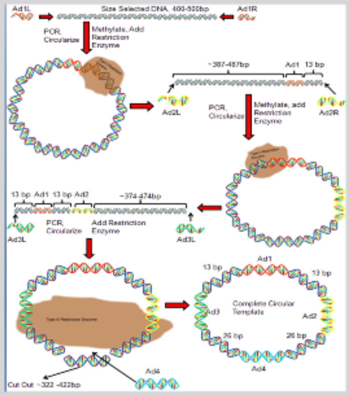


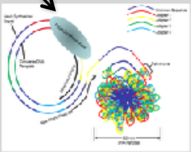

b
Two base encoding: each target nucleotide is interrogated twice
2nd base
1st base
A C G T
Template sequence
ATACAAGA
CGCAGCTC
GCGTGGAG
TATCTTCT
Alignment of colour-space reads to colour-space reference genome
SNP
TCGGATTCAGCCTGCTGCTCTATCA

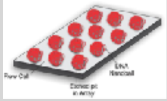
Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46

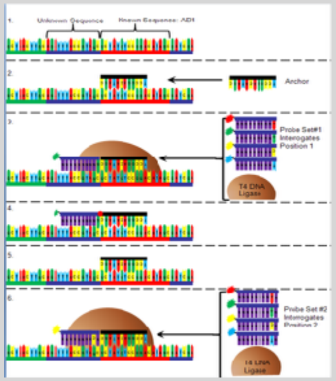
Complete Genomics – Nanoball Sequencing

Has proofreading ability!







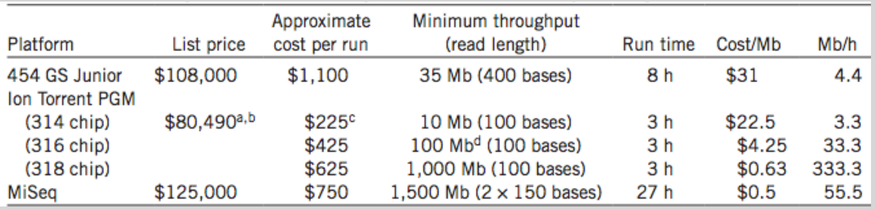


Niedringhaus, TP et al (2011), *Analytical Chem.*, 83, pp. 4327-4341

Wikipedia, “DNA Nanoball Sequencing”, September 26, 2012

“Benchtop” Sequencers

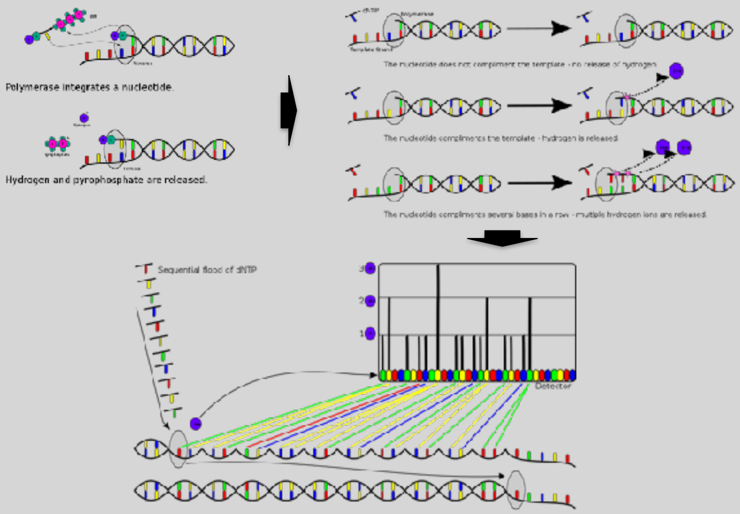
- Lower cost, lower throughput alternative for smaller scale projects
- Currently three significant platforms
  - Roche 454 GS Junior
  - Life Technology Ion Torrent
    - Personal Genome Machine (PGM)
    - Proton
  - Illumina MiSeq



Platform	List price	Approximate cost per run	Minimum throughput (read length)	Run time	Cost/Mb	Mb/h
454 GS Junior	\$108,000	\$1,100	35 Mb (400 bases)	8 h	\$31	4.4
Ion Torrent PGM						
(314 chip)	\$80,490[a,b]	\$225[c]	10 Mb (100 bases)	3 h	\$22.5	3.3
(316 chip)		\$425	100 Mb[d] (100 bases)	3 h	\$4.25	33.3
(318 chip)		\$625	1,000 Mb (100 bases)	3 h	\$0.63	333.3
MiSeq	\$125,000	\$750	1,500 Mb (2 × 150 bases)	27 h	\$0.5	55.5

Loman, NJ (2012), *Nat. Biotech.*, 5, pp. 434-439

PGM - Ion Semiconductor Sequencing



Polymerase integrates a nucleotide.
Hydrogen and pyrophosphate are released.
dNTP
Polymerase
Template Strand
The nucleotide does not compliment the template - no release of hydrogen.
The nucleotide compliments the template - hydrogen is released.
The nucleotide compliments several bases in a row - multiple hydrogen ions are released.
Sequential flood of dNTP
3
2
1
Detector

Wikipedia, “Ion Semiconductor Sequencing”, September 26, 2012