Introduction to sequencing

Benjamin Neale

March 8th, 2011

Massachusetts General Hospital
Harvard Medical School
BROAD INSTITUTE

Direct Sequencing has Enormous Potential…

Ng, Shendure: Miller syndrome, 4 cases

- exome sequenced reveals causal mutations in DHODH

Lifton: Undiagnosed congenital chloride diarrhoea (consanguinous)

- Exome seq reveals homozygous SLC23A chloride ion transporter mutation

Worthey, Dimmock: 4-year old, severe unusual IBD

- exome seq reveals XIAP mutation (at a highly conserved aa)

Jones, Marra: Secondary lung carcinoma unresponsive to erlotinib

- Genome and transcriptome sequencing reveals defects

Mardis, Wilson: acute myelocytic leukaemia but not classical translocation

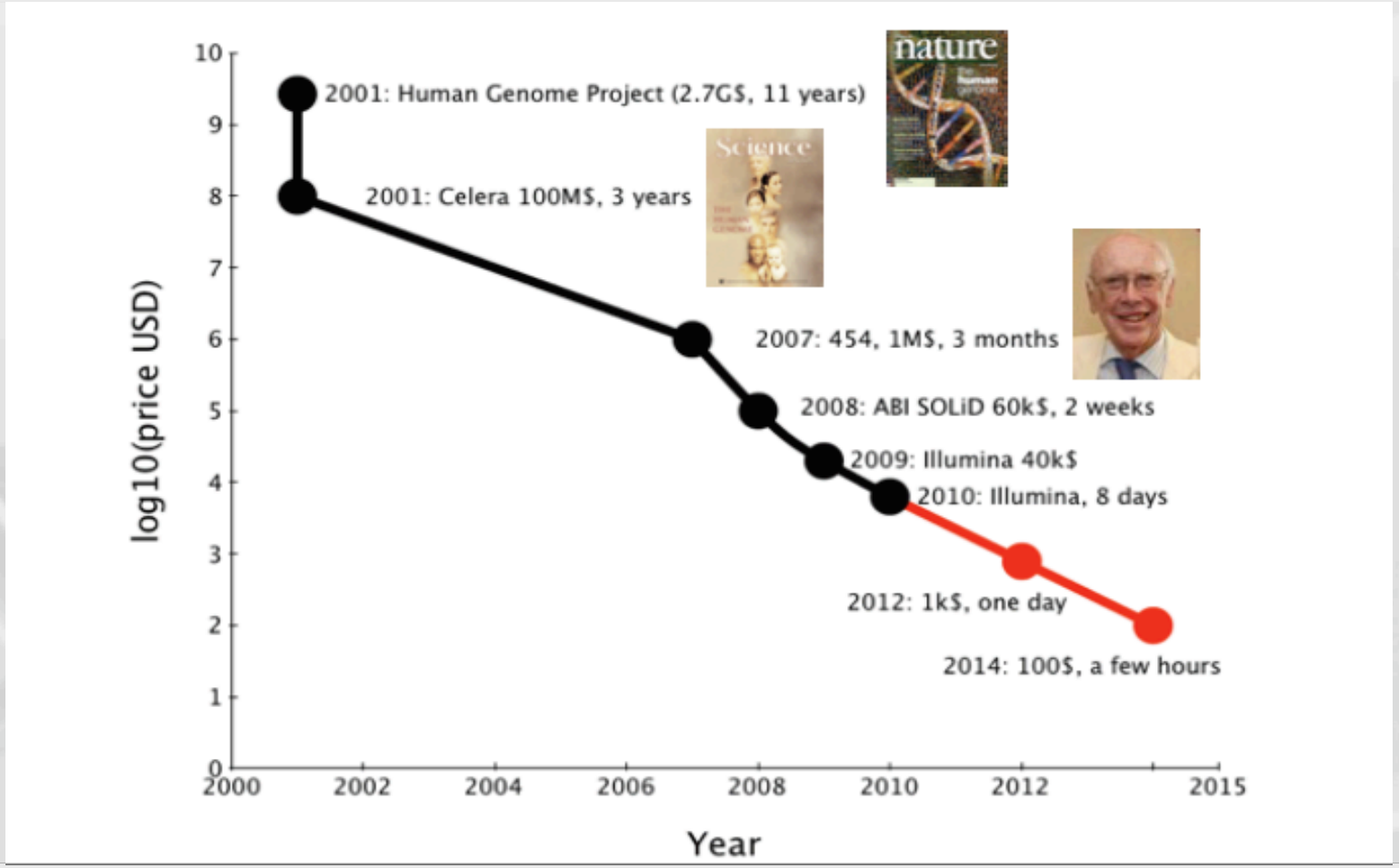
- Genome sequencing (1 week + analysis) reveals PML-RARA translocation

…and tremendous challenges

- Managing and processing vast quantities of data into variation
- Interpreting millions of variants per individual
  - An individual's genome harbors
    - ~80 point nonsense mutations
    - ~100-200 frameshift mutations
    - Tens of splice mutants, CNV induced gene disruptions

For very few of these do we have any conclusive understanding of their medical impact in the population

Cost of Sequencing



So where does this data originate?

Exome vs Genome

Exome

- ~1.5% of genome
- All exons
  - Some pseudogenes missed
- Deeper coverage
  - Average 60-180x
- Useful for Mendelian disease
- Requires library construction

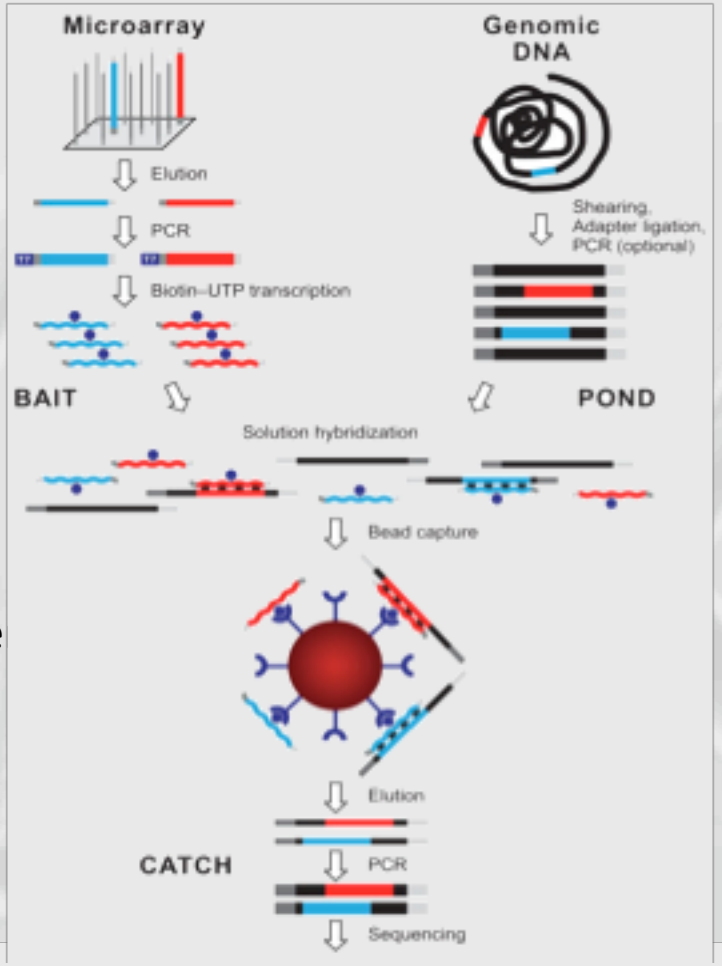
Genome

- Almost entire genome
- Much lower pass
  - Currently 4-12x
- Useful for structural variation/CNVs
- Shotgun sequence
  - Randomly shear and capture

Next Generation Sequencing
Hybrid Selection – Targeted

Steps:

1. Generate pool of oligonucleotides
2. PCR amplify
3. *In vitro* transcription for single strand RNA bait



Steps:

1. Randomly shear DNA
2. Ligate an adapter
3. PCR amplify

1. Bait anneals to pond mate
2. Captured by the magnetic
3. PCR amplified
4. Analyzed on a next-gen sequence machine

Gnirke et al. Nat. Biotechnology, 2009

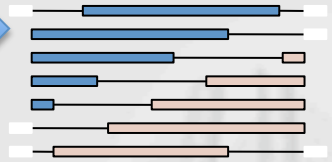
From unmapped reads to true genetic variation in next-generation sequencing data

Solexa

SOLiD

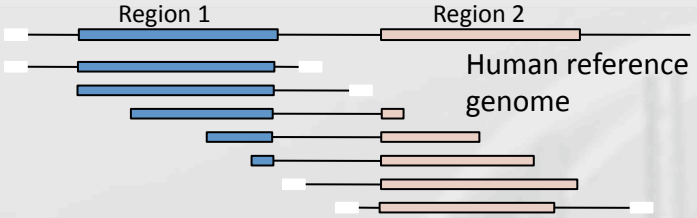
454

Raw short reads



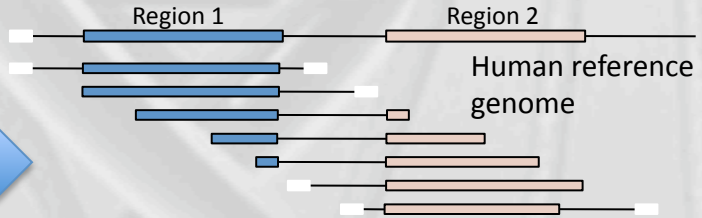
A single run of a sequencer generates ~50M ~75bp short reads for analysis

Mapping and alignment



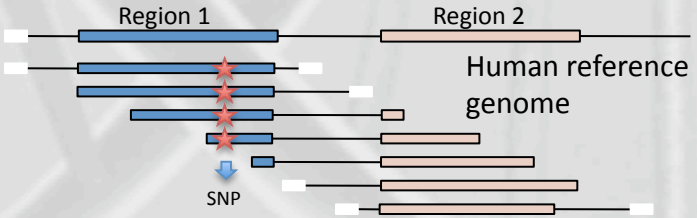
The origin of each read from the human genome sequence is found

Quality calibration and annotation



The quality of each read is calibrated and additional information annotated for downstream analyses

Identifying genetic variation



SNPs and indels from the reference are found where the reads collectively provide evidence of a variant

Core QC concepts

- Coverage
  - # reads at a position
- Transition:Transversion ratio (Ti:Tv or Ts:Tv)
  - Within vs between type (purine or pyrimidine)

	A	C	G	T
A	-	Tv	Ti	Tv
C	Tv	-	Tv	Ti
G	Ti	Tv	-	Tv
T	Tv	Ti	Tv	-

Random Ti:Tv = 0.5 ; Exome Ti:Tv > Genome Ti:Tv (roughly 3.25 and 2.25 respectively)

Core QC Concepts II

- Singleton, doubleton, … n-ton
  - # of copies of non-reference allele
- Case:control counts
  - Consider each set of n-ton
- Mutation types
  - Synonymous/Silent – Same amino acid
  - Missense – Different amino acid
  - Nonsense – Premature stop codon
  - Function – All but synonymous

Core Concepts III

- Read length: # bases assayed by sequencer
  - Varies by technology
  - Shorter the read, harder the alignment
- Mapping Quality Score
  - Measure of how well (uniquely) the read maps
- Base Alignment Quality
  - Probability of the base being misaligned

Session overview

1. Mapping/alignment [Hyun + Goncalo]
2. Indel/CNV calling [Kai]
3. Variant calling + basic QC [Hyun + Goncalo]
4. Rare variant association issues [Ben]
5. PLINK/SEQ practical [Shaun]
6. Bioinformatics/1KG [Goncalo + Hyun]