Introduction to NGS

Buenos Aires, Octubre 2011

Javier Santoyo
jsantoyo@cipf.es
http://bioinfo.cipf.es

Bioinformatics and Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)
Next Generation Sequencing

NGS has brought high speed not only to genome sequencing and personal medicine, but has also change the way we do genome research:

Got a question on genome organization:
SEQUENCE IT!!!!
From genotype to phenotype.

Genes in the DNA...

...code for proteins...

>protein kunase
acctgtgatgacaggactgtgatgatgctgatgatgcatgcatgctgactactgatgtgggggctattga
ttgatgtctatc....

...whose structure accounts for function...

...produces the final phenotype...

...plus the environment...

Data is information
Bioinformatics tools for pre-genomic sequence data analysis

- Information
- Sequence
- Molecular databases
- Search results
- Phylogenetic tree
- Alignment
- Conserved region
- Motif databases
- Motif
- Secondary and tertiary protein structure
The aim:
Extracting as much information as possible for one single data
High Throughput Technologies

• 1988 arrayed DNAs were used
• 1991 oligonucleotides are synthesized on a glass slide through photolithography (Affymax Research Institute)
• 1995 DNA Microarrays
• 1997 Genome wide Yeast Microarray
From genotype to phenotype.

 kırotein kunase
acctttgtggcagggactgtatgc
gatctatgctgatgcatgcatgctgactact
gatgtgggggctattgacttgatgtctatc...

...when expressed in the proper moment and place...

A typical tissue is expressing among 5000 and 10000 genes

Next Generation Sequencing
SOLID 6Gbp per round

...code for proteins...

That undergo post-translational modifications, somatic recombination...

Genes in the DNA...

100K-500K proteins

...whose structures account for function...

Each protein has an average of 8 interactions

Data ≠ information

...whose final effect configures the phenotype...

...which can be different because of the variability.

...conforming complex interaction networks...

10 million SNPs

...in cooperation with other proteins...

(in the functional post-genomics scenario)

...when expressed in the proper moment and place...

...code for proteins...

That undergo post-translational modifications, somatic recombination...
High Throughput Technologies

The road of excess leads to the palace of wisdom.

William Blake (1757 – 1827)
poet, painter, and printmaker
Pre & Post-genomic databases

EMBL database growth (March 2009)
## History of DNA Sequencing

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1870</td>
<td>Miescher: Discovers DNA</td>
</tr>
<tr>
<td>1940</td>
<td>Avery: Proposes DNA as ‘Genetic Material’</td>
</tr>
<tr>
<td>1953</td>
<td>Watson &amp; Crick: Double Helix Structure of DNA</td>
</tr>
<tr>
<td>1965</td>
<td>Holley: Sequences Yeast tRNA(^{\text{Ala}})</td>
</tr>
<tr>
<td>1970</td>
<td>Wu: Sequences (\lambda) Cohesive End DNA</td>
</tr>
<tr>
<td>1977</td>
<td>Sanger: Dideoxy Chain Termination</td>
</tr>
<tr>
<td>1980</td>
<td>Gilbert: Chemical Degradation</td>
</tr>
<tr>
<td>1986</td>
<td>Messing: M13 Cloning</td>
</tr>
<tr>
<td>1990</td>
<td>Hood et al.: Partial Automation</td>
</tr>
<tr>
<td>2002</td>
<td>Cycle Sequencing</td>
</tr>
<tr>
<td>2008</td>
<td>Next Generation Sequencing</td>
</tr>
</tbody>
</table>

**Efficiency (bp/person/year):**

- 1
- 15
- 150
- 1,500
- 15,000
- 25,000
- 50,000
- 200,000
- 50,000,000
- 100,000,000,000

Adapted from Eric Green, NIH; Adapted from Messing & Llaca, *PNAS* (1998)
NGS is here ...

ARTICLES

DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome


Acute myeloid leukaemia is a highly malignant haematopoietic tumour that affects about 13,000 adults and children in the United Kingdom each year.
The size of the human genome is ~ 3 \times 10^9 \text{ bp}; almost all of its complexity is in single-copy DNA.

The human genome is thought to contain ~30,000-40,000 genes.

Computing capabilities (CPU power doubles in ~18-24 months, hard drive capacity doubles in ~12 months, network bandwidth doubles in ~20 months) should increase: \textbf{7-10x} in 5 years.

Data projection in 3-5 years: \textbf{100x} increase in sequencing volume. Still new technologies with higher throughput to come very soon!!!
Next generation technology is here

Relative throughput of the different technologies. NGS emerges with a potential of data production that will, eventually wipe out conventional HT technologies in the years coming.
Some of the most common applications of NGS

**RNA-seq**
- Transcriptomics
  - Quantitative
  - Descriptive
  - alternative splicing
  - miRNA profiling

**ChIP-seq**
- Protein-DNA interactions
- Active transcription factor binding sites

**Resequencing**
- Mutation calling
- Profiling

**De novo sequencing**

**Copy number variation**

**Metagenomics**

**Metatranscriptomics**
Clone the DNA.

Generate a ladder of labeled (colored) molecules that are different by 1 nucleotide.

Separate mixture on some matrix.

Detect fluoroscope by laser.

Interpret peaks as string of DNA.

Strings are 500 to 1,000 letters long

1 machine generates 57,000 nucleotides/run

Assemble all strings into a genome.
Basics of the “new” technology

- Get DNA.
- Attach it to something.
- Extend and amplify signal with some color scheme.
- Detect fluorochrome by microscopy.
- Interpret series of spots as short strings of DNA.
- Strings are 30-500 letters long
- Multiple images are interpreted as 0.4 to 1.2 GB/run/day (1,200,000,000 letters/day).
- Map or align strings to one or many genome.
<table>
<thead>
<tr>
<th></th>
<th><strong>Sanger (1st-gen) Sequencing</strong></th>
<th><strong>Next-Gen Sequencing, and 3rd generation</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Genome</td>
<td>Human (early drafts), model organisms, bacteria, viruses and mitochondria (chloroplast), low coverage</td>
<td>New human (!), individual genome, exomes, 2,500 normal (1K genome project), 25,000 cancer (TCGA and ICGC initiatives), CNV, matched control pairs, time course, rare-samples</td>
</tr>
<tr>
<td>RNA</td>
<td>cDNA clones, ESTs, Full Length Insert cDNAs, other RNAs</td>
<td>RNA-Seq: Digitization of transcriptome, alternative splicing events, miRNA, allele specific transcripts</td>
</tr>
<tr>
<td>Communities</td>
<td>Environmental sampling, 16S RNA populations, ocean sampling,</td>
<td>Human microbiome, deep environmental sequencing, Bar-Seq</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>Epigenome, rearrangements, ChIP-Seq</td>
</tr>
</tbody>
</table>
NGS technologies

Cost-effective
Fast
Ultra-throughput
Cloning-free
Short reads
Differences between the various platforms:

- Nanotechnology used.
- Resolution of the image analysis.
- Chemistry and enzymology.
- Signal to noise detection in the software.
- Software/images/file size/pipeline.
- Cost.
General ways of dealing at the sequences:

- Assemble them and look at what you have
- You map them (align against a known genome) and then look at what you have.
- Or a mixture of both!

- Sometimes you select the DNA you are sequencing
- or you try to sequence everything

- Depends on biological question, sequencing machine you have, and how much time and money you have
Next-gen sequencers

AB/SOLiDv3, Illumina/GAII short-read sequencers
(10+Gb in 50-100 bp reads, >100M reads, 4-8 days)

454 GS FLX pyrosequencer
(100-500 Mb in 100-400 bp reads, 0.5-1M reads, 5-10 hours)

ABI capillary sequencer
(0.04-0.08 Mb in 450-800 bp reads, 96 reads, 1-3 hours)

From John McPherson, OICR
2009/10

bases per machine run

- 100 Gb
- 10 Gb
- 1 Gb
- 100 Mb
- 10 Mb
- 1 Mb

read length

- 10 bp
- 100 bp
- 1,000 bp

- AB SOLiDv3
  120Gb, 100 bp reads

- Illumina HiSeq
  100Gb, 150bp reads

- 454 GS FLX Titanium
  0.4-0.6 Gb, 100-400 bp reads

- ABI capillary sequencer
  (0.04-0.08 Mb, 450-800 bp reads)

Adapted from John McPherson, OICR
Next Generation Methods

• 3 main platforms:
  • Solexa/illumina
  • Roche 454
  • ABI SOLiD

• Follow an approach similar to Sanger sequencing, but do away with separation of fragments by size and “read” the sequence as the reaction occurs

• Several different “next generation” sequencing platforms developed and commercialized, more on the way.

• Simultaneously sequence entire libraries of DNA sequence fragments
Table 1 | Comparison of next-generation sequencing platforms

<table>
<thead>
<tr>
<th>Platform</th>
<th>Library/template preparation</th>
<th>NGS chemistry</th>
<th>Read length (bases)</th>
<th>Run time (days)</th>
<th>Gb per run</th>
<th>Machine cost (US$)</th>
<th>Pros</th>
<th>Cons</th>
<th>Biological applications</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche/454's GS FLX Titanium</td>
<td>Frag, MP/emPCR</td>
<td>PS</td>
<td>330*</td>
<td>0.35</td>
<td>0.45</td>
<td>500,000</td>
<td>Longer reads improve mapping in repetitive regions; fast run times</td>
<td>High reagent cost; high error rates in homopolymer repeats</td>
<td>Bacterial and insect genome de novo assemblies; medium scale (&lt;3 Mb) exome capture; 16S in metagenomics</td>
<td>D. Muzny, pers. comm.</td>
</tr>
<tr>
<td>Illumina/ Solexa's GA,</td>
<td>Frag, MP/solid-phase</td>
<td>RTs</td>
<td>75 or 100</td>
<td>4±, 9±</td>
<td>18±, 35±</td>
<td>540,000</td>
<td>Currently the most widely used platform in the field</td>
<td>Low multiplexing capability of samples</td>
<td>Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics</td>
<td>D. Muzny, pers. comm.</td>
</tr>
<tr>
<td>Life/AGI's SOLiD 3</td>
<td>Frag, MP/emPCR</td>
<td>Cleavable probe SBL</td>
<td>50</td>
<td>7±, 14±</td>
<td>30±, 50±</td>
<td>595,000</td>
<td>Two-base encoding provides inherent error correction</td>
<td>Long run times</td>
<td>Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics</td>
<td>D. Muzny, pers. comm.</td>
</tr>
</tbody>
</table>

*Average read-lengths. ±Fragment run. ±Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.
**454 (Roche)**

- First next generation method to be commercially available

- Uses a “sequencing by synthesis” approach:
  - DNA is broken into pieces of 500-1,000 bp, ligated to adaptors, and amplified on tiny beads by PCR (emulsion PCR)
  - Beads (with DNA attached) are placed into tiny wells (one bead per well) on a PicoTiterPlate that has millions of wells. Each well is connected to an optical fiber.
  - DNA is sequenced by adding polymerase and DNA bases containing pyrophosphate. The different bases (A,C,G,T) are added sequentially in a flow chamber
  - When a base complementary to the template is added, the pyrophosphate is released and a burst of light is produced
  - The light is detected and used to call the base

- Initially 100-150 bp, but they have been improved to 400-500

- $>1$ million, filter-passed reads per run (10 hours)

- 1 billion bases per day
Roche 454 pyrosequencing

**Principle**
Preparation of the DNA includes: DNA fragmentation (nebulization), DNA size selection, Fragment end polishing, Adaptor ligation, Library immobilization, fill in reaction and ssDNA library isolation. At the end of these steps, the DNA fragments are ready for the emulsion PCR (emPCR).

**emPCR** include the immobilisation of the DNA fragments on capture beads (1 molecule / bead), emulsification (1 bead / aqueous microreactor), amplification and indirect enrichment resulting in an immobilized and amplified library.

**Sequencing** includes a prewash, the loading DNA library beads, enzyme beads (PPIase) and packing beads on the picotiter plate (PTP). Run over night. At the end of these steps you get your data.
Roche 454 pyrosequencing

Figure 1 | The principle of Pyrosequencing and the output Pyrogram™. Double peak heights indicate incorporations of two nucleotides in a row.
Roche / 454 : GS FLX

- Made for de novo sequencing (longer reads).

- Too expensive for resequencing.

- New bacterial genomes.

- Pyrosequencing. Bias with long polinucleotide stretches
Roche 454

<table>
<thead>
<tr>
<th><strong>Throughput</strong></th>
<th>400-600 million high-quality, filter-passed bases per run* 1 billion bases per day</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Run Time</strong></td>
<td>10 hours</td>
</tr>
<tr>
<td><strong>Read Length</strong></td>
<td>Average length = 400 bases</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>Q20 read length of 400 bases (99% at 400 bases and higher for prior bases)</td>
</tr>
<tr>
<td><strong>Reads per run</strong></td>
<td>&gt;1 million high-quality reads</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>Trace data accepted by NCBI since 2005</td>
</tr>
<tr>
<td><strong>Computing Requirements</strong></td>
<td>Cluster recommended (Roche GS FLX Titanium Cluster available)</td>
</tr>
<tr>
<td><strong>Robustness</strong></td>
<td>No complex optics or lasers; reagents have long shelf life</td>
</tr>
</tbody>
</table>
GS Junior, benchtop

**System Performance**

<table>
<thead>
<tr>
<th><strong>Throughput</strong></th>
<th>35 million high-quality, filtered bases per run*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Run Time</strong></td>
<td>10 hours sequencing</td>
</tr>
<tr>
<td></td>
<td>2 hours data processing</td>
</tr>
<tr>
<td><strong>Avg. Read Length</strong></td>
<td>400 bases*</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>Q20 read length of 400 bases</td>
</tr>
<tr>
<td></td>
<td>(99% accuracy at 400 bases)</td>
</tr>
<tr>
<td><strong>Reads per Run</strong></td>
<td>100,000 shotgun, 70,000 amplicon</td>
</tr>
<tr>
<td><strong>Sample Input</strong></td>
<td>gDNA, amplicons, cDNA, or BACs depending on the application</td>
</tr>
<tr>
<td><strong>Physical Dimensions</strong></td>
<td>40 cm wide x 60 cm deep x 40 cm high (the size of a laser printer)</td>
</tr>
<tr>
<td></td>
<td>Weight = 55 lbs.</td>
</tr>
<tr>
<td><strong>Computing</strong></td>
<td>Linux-based OS on HP desktop computer included.</td>
</tr>
<tr>
<td></td>
<td>All software is point-and-click.</td>
</tr>
</tbody>
</table>

*Typical results.* Average read length and number of reads depend on specific sample and genomic characteristics.
Solexa (Illumina)

- Uses a “sequencing by synthesis” approach:
  - DNA is broken into small fragments and ligated to an adaptor.
  - The fragments are attached to the surface of a flow cell and amplified.
  - DNA is sequenced by adding polymerase and labeled reversible terminator nucleotides (each base with a different color).
  - The incorporated base is determined by fluorescence.
  - The fluorescent label is removed from the terminator and the 3’ OH is unblocked, allowing a new base to be incorporated.

- Started with 35 bp, increased now to up to 150 bp
- One run can give up to 640 million paired-end reads
- Up to 6.5 Gb per day for a 2 x 100 bp run
Solexa / illumina

**Illumina/Solexa**

**Solid-phase amplification**

One DNA molecule per cluster

- **Template dNTPs and polymerase**
- **Sample preparation DNA (5 µg)**
- **Cluster growth**
- **100–200 million molecular clusters**
- **Bridge amplification**

**Solexa / illumina**

**a** Illumina/Solexa — Reversible terminators

- Incorporate all four nucleotides, each label with a different dye
- Wash, four-colour imaging
- Cleave dye and terminating groups, wash

**b**

- Top: CATCGT
- Bottom: CCCCCC

Repeat cycles
Solexa / illumina

Base calling from raw data

The identity of each base of a cluster is read off from sequential images
Solexa-HiSeq 2000

200 Gb/run in 8 days
2x100 bp fragments
2 billion reads per run
SOLiD (ABI)

“sequencing by ligation” method
 Does not use polymerase, instead uses DNA ligase for sequencing:
  - DNA is broken into small fragments and ligated to an adaptor.
  - The fragments are attached to beads and amplified by emulsion PCR. Beads are attached to the surface of a glass slide.
  - DNA is sequenced by adding 8-mer fluorescently labeled oligonucleotides
  - If an oligo is complementary to the template, it will be ligated and 2 of the bases can be called.
  - The attached oligo is then cut to remove the label and the next set of labeled oligos are added
  - The process is repeated from different starting points (using different universal primers) so that each base is called twice

Colors
- Colorspace
- 25bp-50bp
- 85 million reads per run
- 1 Gb per day
SOLiD
SOLiD

* Sequencing output in “color space”

** Needs reference genome to translate to base space.
(a) Solid sequencing process

Round

1
2
3
4
5

Ligation cycle

1 2 3 4 5

(b) Principles of two base encoding

Reference: A C G G T C G T C G T C G T C G T

2 base probes

Sequence 1

Sequence 2

Sequence 3

Sequence 4

Sequence 5

Wild type

2 color change = SNP

1 color change = error

Incorrect color change = error

1 bp deletion
Use the following steps to encode a DNA sequence ATCAAGCCTC*:

1. start at the 5' end,

2. replace the di-base AT at this position with its corresponding code 3 from the table,

3. advance by one base, which exposes the TC di-base, and

4. continue, as shown below.

Base Sequence: A T C A A G C C T C
Color String: 3 2 1 0 2 3 0 2 2
Union of Biochemistry) codes. So let $B = \{A, C, G, T\}$. The color code should satisfy the following requirements:

For all bases $b, d, e$ in $B$:
1. The available colors are 0, 1, 2, and 3:
   \[ \text{color } (bd) \in \{0, 1, 2, 3\}. \]

2. Two different di-bases that have the same first base get different colors: \( \text{color } (bd) \neq \text{color } (be) \) if \( d \neq e \).
   For example, \( \text{color } (AC) \neq \text{color } (AG) \).

3. A di-base and its reverse get the same color:
   \[ \text{color } (bd) = \text{color } (db). \]
   For example, \( \text{color } (AC) = \text{color } (CA) \).

4. Monodibases get the same color:
   \[ \text{color } (bb) = \text{color } (dd). \]

\[
\begin{array}{c|cccc}
\text{First Base} & A & C & G & T \\
\hline
A & 0 & 1 & 2 & 3 \\
C & 1 & 0 & 3 & 2 \\
G & 2 & 3 & 0 & 1 \\
T & 3 & 2 & 1 & 0 \\
\end{array}
\]
SOLiD color space

That is, color \((AA) = color \((CC) = color \((GG) = color \((TT)\). The following are not requirements, but interesting properties that follow from these four. Property 5 follows from requirements 2 and 3, and will make our construction easier.

<table>
<thead>
<tr>
<th>First Base</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Panel E
5. Two different di-bases that nevertheless have the same second base get different colors:
   color \((bd)\) \(\neq\) color \((cd)\), if \(b \neq c\).

For example, color \((AC)\) \(\neq\) color \((TC)\). Property 6 also follows from requirements 1-4, but it is most easily verified against the completed code (Figure 3, Panel E).

6. A di-base and its complement get the same color:
   color \((b^C d^C)\) = color \((d^C b^C)\).
   For example, color \((AC)\) = color \((TG)\).
SOLiD color space
Its format is

>`TAG_ID
Color_space`

e.g.

>`1_88_1830_R3
G32113123201300232320`  
>`1_89_1562_R3
G231331312333333101320`
AB SOLiD: Dibase Sequencing

AB SOLiD reads look like this:

T012233102
TGAGCGTTC

T012033102
TGAATAGGA

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
AB SOLiD: Variations

Figure 5. Examples of polymorphisms in color space.
AB SOLiD: Valid substitutions

Single SNP Sites

If 2 colors present: e.g.
- Reverse the colors
- Use other 2 colors, both combinations

If 1 color present: e.g.
- Can be the other 3 color pairs:

Example using rules above

Valid 2 color change
- C T T
- C G T
- C A T

Invalid 1 color change
- T T T
- A A T
- G G T
- T C C

Invalid 2 color change
- C G C
- C C G
- C T A
- C A G
- T G T
- G T T
5500 xl-u SOLiD

- 180 Gb/run (microbeads)
- 300 Gb/run (nanobeads)
- 35-75 bp fragments
- 2.8 - 4.8 billion reads/run
- 2x6 lanes/run
- 96 bar-codes
- 99.99% accuracy
SOLiD 5500

* Fifth 3-based encoded primer
* Sequencing output in base space
* No reference needed
Helicos

a. Illumina/Solexa — Reversible terminators

- Incorporate all four nucleotides, each label with a different dye
- Wash, four-colour imaging
- Cleave dye and terminating groups, wash
- Repeat cycles

b. Example images of sequencing data

<table>
<thead>
<tr>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top: CATCGT</td>
<td>Bottom: CCCCCC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

c. Helicos BioSciences — Reversible terminators

- Incorporate single, dye-labelled nucleotides
- Wash, one-colour imaging
- Cleave dye and inhibiting groups, cap, wash
- Repeat cycles

d. Example images of sequencing data

<table>
<thead>
<tr>
<th>C</th>
<th>T</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top: CTAGTG</td>
<td>Bottom: CAGCTA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PacBio

Real time DNA synthesis
Up to 12000 nt??
50 bases/second??
Real-Time DNA Sequencing from Single Polymerase Molecules

John Eid,* Adrian Fehr,† Jeremy Gray,* Khai Luong,* John Lyle,* Geoff Otto,* Paul Peluso,* David Rank,† Primo Baybayan, Brad Bettman, Arkadiusz Bujil, Keith Bjornson, Bithan Chaudhuri, Frederick Christians, Ronald Ciero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Garretson, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xianguo Kong, Ronald Kus, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Macham, Devon Murphy, Imre Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorensen, Austin Tomney, Kevin Travers, Mark Trask, John Vincelli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarini, Peter Zhao, Frank Zhong, Jonas Koffler,* Stephen Turner†

Previously been shown in both birds and butterflies that structural color can enhance pigment color either by an additive or a contrast effect [8, 16, 29, 30]. This interplay of structure and pigment may therefore also add to the diversity of pollination cues utilized by the flowers of many angiosperm species.

References and Notes
31. W. H. F. H. Kebi, P. Kebi, R. Sura, P. Costa, and I. Barmburg for discussions and suggestions for the manuscript. Funded by Natural Environmen Research Council grant NERC/0055531, Engineering and Physical Sciences Research Council grant EPSRC/0881401, the European R&D-30-5 network Pattern, the Cambridge University Research Exchange, and German Academic Exchange Service DNA.

Supporting Online Material
www.sciencemag.org/cgi/content/full/323/5910/130/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 and S2
References
22 September 2008; accepted 8 November 2008
10.1126/science.1164256
Ion Torrent

- $50,000
- $500/sample
- 1 hour/run
- >200 nt lengths
- Reads H+ released by DNA polymerase
## Comparison

<table>
<thead>
<tr>
<th><strong>Roche 454</strong></th>
<th><strong>Solexa</strong></th>
<th><strong>SOLiD</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Long fragments</td>
<td>• Short fragments</td>
<td>• Short fragments</td>
</tr>
<tr>
<td>• Errors: poly nts</td>
<td>• Errors: Hexamer bias</td>
<td>• Color-space</td>
</tr>
<tr>
<td>• Low throughput</td>
<td>• High throughput</td>
<td>• High throughput</td>
</tr>
<tr>
<td>• Expensive</td>
<td>• Cheap</td>
<td>• Cheap</td>
</tr>
<tr>
<td>• De novo sequencing:</td>
<td>• Resequencing:</td>
<td>• Resequencing:</td>
</tr>
<tr>
<td>• Amplicon sequencing</td>
<td>• ChipSeq</td>
<td>• ChipSeq</td>
</tr>
<tr>
<td></td>
<td>• RNASeq</td>
<td>• RNASeq</td>
</tr>
<tr>
<td></td>
<td>• MethylSeq</td>
<td>• MethylSeq</td>
</tr>
</tbody>
</table>
Applications

- De novo sequencing
- Resequencing
- Exome Sequencing
- RNA-seq
- Genome annotation
- Chip-seq
- Methyl-seq
- ..........
Applications

- De novo sequencing
- Resequencing
- Exome Sequencing
- RNA-seq
- Genome annotation
- Chip-seq
- Methyl-seq
- .......
Basic steps NGS data processing

QC and read cleaning
Basic steps NGS data processing

QC and read cleaning

Mapping
Basic steps NGS data processing

- QC and read cleaning
- Mapping
- Feature identification
Basic steps NGS data processing

- QC and read cleaning
- Mapping
- Feature identification
- SNVs
  - Indels
  - Rearrangements
- RPKM
- Splicing
- DNA
- Binding site
RNA-seq

Elucidate gene models

Quantify gene expression
RNA-seq

Elucidate gene models
**RNA-seq protocol**

**Total RNA**

1. **2nd strand synthesis**

2. **1st strand synthesis**

3. **mRNA preparation**
   - oligodT
   - RiboZ

4. **Fragmentation**

*Solexa Pair-End*
RNA-seq protocol (II)

- Adenylation 3’ ends
- Ligate adapters
- Amplification

SEQUENCING!
Strand-specific RNAseq

dUTP library

1. Purification of mRNA
2. Fragmentation
3. First-strand synthesis
4. Second-strand synthesis
   - dTTP replaced by dUTP
5. End Repair
6. Adenylate 3’ ends
7. Ligate adapters

Diagram shows the process of strand-specific RNA sequencing, starting with the purification of mRNA, followed by fragmentation, first-strand synthesis, second-strand synthesis with dTTP replaced by dUTP, end repair, adenylate 3’ ends, and ligation of adapters.
Strand-specific RNA-seq

Purification of cDNA (library size-selection)

Treatment with uracil-N glycosilase (digests dUTPs)

180-480bp cut – adapter size = 122-422bp insert size

Amplification of the original strand
File formats

**fastq: sequence data and qualities**

**SAM/BAM: mapping data and qualities**
Some Figures

How much does it “cost” (computationally) to sequence a human transcriptome?

<table>
<thead>
<tr>
<th>One human transcriptome: 100 Million reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Solexa run ==8 lanes ==25 M reads/lane==2 x 4 G fastq/lane (PE)</td>
</tr>
<tr>
<td>space</td>
</tr>
<tr>
<td>Mapping @ processor 12 cores, 48 GB RAM , 4TB disk</td>
</tr>
<tr>
<td>SAM (Ascii) / BAM (Binary) output</td>
</tr>
<tr>
<td>32 Gb disk</td>
</tr>
<tr>
<td>24 hours</td>
</tr>
<tr>
<td>36 Gb / 9 Gb</td>
</tr>
</tbody>
</table>
Applications of RNAseq

Qualitative:
* Alternative splicing
* Antisense expression
* Extragenic expression
* Alternative 5' and 3' usage
* Detection of fusion transcripts

Quantitative:
* Differential expression
* Dynamic range of gene expression

Tools:
- Tophat/Cufflinks
- Scripture
- Alexa
- edgeR
- DESeq
- baySeq
- NOISeq
Advantages of RNAseq?

- Non targeted transcript detection
- No need of reference genome
- Strand specificity
- Find novels splicing sites
- Larger dynamic range
- Detects expression and SNVs
- Detects rare transcripts

and.... are there any disadvantages?????
Resequencing
Exome Sequencing

1. Produce shotgun library
2. Capture exon sequences
3. Wash & Sequence
4. Map against reference genome
5. Determine variants, Filter, compare patients

Gene A
Gene B
DNA (patient)

candidate genes
Exome capture
The principle: comparison of patients

- Patient 1
- Patient 2
- Patient 3
- Patient 4
- Patient 5
- Patient 6

Validate gene (shares mutation for all patients)
ChipSeq

1. Cell Nucleus

2. Crosslink Protein and Shear DNA

3. Add Protein-Specific Antibody

4. Immunoprecipitate and purify complexes

5. Reverse Crosslinks, Purify DNA and prepare for sequencing

6. Sequence DNA fragment and map to genome
MethylSeq
Census NGS methods

ChIP-Seq
- 1 kb at the SRF locus
  - Serum response factor
  - MyoD factor
  - Binding motifs

mRNA-Seq
- 6 kb at the cdk2 locus
  - Spliced reads
  - Unspliced reads

Methyl-Seq
- 60 kb at the PRDM2 locus
  - Undigested control DNA
  - CpG islands
  - PDRM2 cDNA a
  - PDRM2 cDNA b

Other input preparations
- Chromatin immunoprecipitation
- mRNA extraction
- Methyl-sensitive DNA preparation
- Other (microRNA, 3C, ribonucleoprotein, DNase-hypersensitive sites, nucleosome position, etc.)

Ultrahigh-throughput sequencing