DNA, the stuff of life.

We can think of DNA as a biochemical entity or as a data string. With the advent of high throughput sequencing and an increasingly central role for bioinformatics <http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html>, the tendency is to think of DNA as a string of A’s, C’s, T’s and G’s. However, in seeking to translate information from genome sequencing projects into applied outcomes we need to revisit DNA as a biochemical entity in order to design appropriate assays to detect genetic variation.

The central theme of molecular biology is that DNA makes RNA, RNA makes protein, and these proteins are responsible for phenotypes. An understanding of the variation in DNA can therefore lead to a mechanistic understanding of phenotypic variation. We now know that the central dogma, DNA makes RNA and RNA makes protein is an oversimplification. The ability of RNA to make DNA, the role of small RNA molecules in regulating genes and conditioning genome imprinting has lead to the RNA World hypothesis <http://en.wikipedia.org/wiki/RNA_world_hypothesis>. Despite revisions in our thinking about the central role of DNA, there is still a central assumption that most phenotypic variation can be linked to variation in DNA sequence. DNA is composed of four bases, adenine, guanine, cytosine and thymine with uracil replacing thymine in DNA that has been transcribed into mRNA (Figure 1). In DNA, these bases are attached to a phosphate-deoxyrobose backbone and they pair, adenine with thymine and guanine with cystine, in an anti-parallel manner (Figure 2). The implications of this pairing are that one strand forms a template for the second strand.

![Figure 1. base_num2.jpg](http://www.imb-jena.de/IMAGE.html)

Blackburn and Gait, *Nucleic acids in chemistry and biology*, Oxford University Press New York 1996. structures of the five major purine (R) and pyrimidine (Y) bases of nucleic acids in their dominant tautomeric forms and with the IUPAC numbering system. Image library of Biological Macromolecules: http://www.imb-jena.de/IMAGE.html
**Figure 2.** Chemical structure of DNA, with colored label identifying the four bases as well as the phosphate and deoxyribose components of the backbone.


Credit: Madeleine Price Ball. *Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2*

*This file is licensed under the Creative Commons Attribution ShareAlike 3.0*
Information in the DNA code:

The biochemical structure of DNA determines much of how we interpret a string of sequence. For example, 447 bases from the 5’ region of the tomato PSY1 gene can be depicted as a string as shown below. This sequence is in FASTA format where the symbol ‘>’ denotes a name. The name can be simple, as in the sequence depicted here, or more complex. For example, the NCBI report for this sequence found at <http://www.ncbi.nlm.nih.gov/nuccore/EF534740.1?report=fasta&log$=seqview> lists the name as “>gi|155965506|gb|EF534740.1| Solanum lycopersicum cultivar Red Setter phytoene synthase 1 (psy1) gene, complete cds”. In the FASTA format the name is followed by a paragraph break, and subsequently by a string of DNA or protein sequence.

>PSY1
TATTCTCTAGGGAGATCTACTAGGATATTTATTTCTATAAAACTAGTAAAGTTGGAGGTTGACAAAAAGAAACAAAAATCTTGAAATTGTGTTTACACCAACACGGTCTCTGCTCAATGAGGTTGTTGCTTCTGGACAGTCCTCAAAATGGGACAAGTTTCATGGAATCAGTCCGGGAGGGAAACCGTTTTTTTGATTCATCGAGGCATAGGAATTTGGTGTCCAAATGAGAGAATCAATAGAGGTGGTGGAAAGCAAACTAATAATGGACGGAAATTTTCTGTACGGTCTGCTATTTTGGCTACTCCATCTGGAGAACGGACGATGACATCGGAACAGATGGTCTATGGTGGTTTTGAGGCAGGCAGCCTTGGTGAAGAGGCAACTGAGTCTACAAATGAGGATAGAAGCCCGAT

This DNA string contains all of the information that we need to determine the reverse strand and therefore the six possible open reading frames that might code for a protein (Figure 3).

Figure 3. Screenshot of possible Open Reading Frames identified in 447 bases derived from the 5’ end of the PSY1 gene of tomato. The figure is generated from the NCBI ‘s ‘ORF’ finder: http://www.ncbi.nlm.nih.gov/gorf/gorf.html
Assaying Sequence Variation: Designing rapid indirect tests for sequence variation

By comparing DNA sequences between two or more plant varieties we can identify differences that might affect the protein and therefore the phenotype. Two important points: (1) detecting differences relies on comparative biology; (2) we often make assumptions about what sort of variation will result in a phenotype. These assumptions require some critical inspection. An alignment of the PSY1 sequence from the tomato genotype ‘Red Setter’ with VRT-32-1 reveals two genetic differences: an A > T change at position 69 and an A > G mutation at position 85 (Figure 4)

Figure 4. Alignment of a portion of the tomato PSY1 gene between the Query sequence for ‘Red Setter’ and Subject sequence for VRT-32-1. The alignment was produced using the NCBI align2seq function in BLAST (see: http://blast.ncbi.nlm.nih.gov/)

Designing Assays to detect DNA polymorphisms (differences in sequence) depends on the nature of the mutation. A review of both the types of mutations that occur and detection systems is therefore appropriate. The A > T change at position 69 and the A > G change at position 85 in the alignment illustrated in Figure 4 are examples of Single Nucleotide Polymorphisms (SNPs). These mutations do not affect the length of a defined fragment, they affect the base composition. Another type of mutation is the insertion/deletion class of polymorphism. These mutations can be detected as fragment-length differences.

A special class of insertion/deletion polymorphisms are the Simple Sequence Repeat (SSR) polymorphisms, also called microsatellites <http://en.wikipedia.org/wiki/Microsatellite>. These polymorphisms occur in shot repeat sequences of repeated DNA. For example one variety may have the sequence AGGAGCAACAACACGCTTTA which has three repeats of CAA (or AAC, depending where you start counting from). A second variety may have the sequence AGGAGCAACAACACACGCTTTA where the trinucleotide CAA is repeated five times. Microsatellites typically consist of di-, tri- or tetra-nucleotide repeats. They are highly variable in genomes, with mutation rates that are 5-10 fold higher than SNPs. The increased mutation rate is thought to occur due to slipped strand mis-parring (slippage) during DNA replication. These mutations therefore provide a ‘marker’ that is highly variable, dispersed throughout the genome, and easily detected as size polymorphisms as long as the detection method can distinguish -2, -3, or -4 base differences.

One means of detecting a SNP is to convert it into a size polymorphism. This can be done through digestion with a restriction endonuclease <http://en.wikipedia.org/wiki/Restriction_enzyme>. Restriction endonucleases (RE) are proteins that recognize a specific sequence of DNA, and cleave at or near the sequence. These enzymes are part of the bacterial defense system, serving to protect from external DNA such as that which might arise from viral infection. Thus identifying differences between PCR amplified fragments from two different varieties can be visualized as a fragment difference post digestion with an RE that cuts the amplicon from one variety but not a the second variety. For example the A>G SNP in the PSY1 gene from tomato can be detected between Red Setter and VRT-32-1 by digestion with PshA1 (Figure 5). A SNP polymorphism detected in this way is often referred to as a Cleaved Amplified Polymorphism or CAP (see CAP tutorial).
Figure 5. Differences in restriction endonuclease recognition sites between ‘Red Setter’ (A) and VRT-32-1 (B). The recognition site for PshAI (GACNNNNGTC) is created by the A > G mutation.

Platforms for detecting differences

Size PM
- Gel
  - Agarose
  - Polyacrylamide
  - Stained
  - IR detection
- Capillary Fluorescence detection (ABI) or IR detection (Beckman)

SNP
- Conversion to size PM (CAP)
- ASPE detected using the Luminex or Illumina BeadXpress
- Golden Gate detected using Illumina BeadXpress

Fragment Analysis to Detect Size PM

An example of fragment analysis using a capillary instrument we discuss the Beckman CEQ8800. The CEQ8800 is an 8 capillary instrument with non proprietary software for DNA sequencing, STR (Short Tandem Repeat), and AFLP (Amplified restriction Fragment Length Polymorphic) fingerprinting. The CEQ8800 has diode lasers that excites infrared dyes, and has four detection channels. For fragment analysis, alleles in each sample can be labeled using PCR amplification using primers that are tagged with an IR dye. Four IR dyes are used, three separate dyes for multi-plexing, and the fourth dye is used for the size standard. Genotyping data are analyzed with the CEQ800 Series Fragment Analysis Software that estimates DNA fragment sizes and amounts and identifies alleles represented by specific DNA fragments. This software runs on Windows2000 and WindowsXP professional.

An example of a gel-based fragment analysis system is the LICOR IR2 DNA analyzer. Also developed as a sequencer, the IR2 uses a polyacrylamide slab-gel to separate fragments based on size. Detection is based on IR dyes, with two dyes available (700 nm and 800 nm channels). Primers are synthesized directly with the label, or indirectly using an M13 tail. <protocol>
Scoring data for INDEL and SSR Polymorphisms

Figure 6. Screen Shot showing the Beckman CEQ8800 output for one capillary. The red dye is the size marker, blue and green dyes are used to label PCR products. This capillary is separating three fragments.

Figure 7. Screen shot of fragment quantification for genotyping using the Beckman CEQ8800.
Fragment Analysis on the LICOR IR2.

Like the Beckman CEQ8800, the LICOR IR2 also uses IR dyes. In contrast, the LICOR separates fragments on a traditional slab-gel, detects the IR signal, and then creates a ‘pseudo gel image’ for display.

**Figure 8.** A series of pseudo gel images from the LiCOR IR2. (A) displays a simple bi-allelic pattern that is easy to interpret and score. (B) shows a more complex pattern that is difficult to interpret in terms of alleles. (C) shows an SSR marker that is showing slippage.

**Scoring exercise:**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>31</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>H</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>H</td>
<td>H</td>
<td>A</td>
<td>H</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the pseudo gel images, each lane represents a single individual. Small fragments travel through the polyacrylamide gel faster than large fragments. The LICOR IR2 allows us to load 64 individuals (including size markers). Markers may be multiplexed based on size and on the IR 700 or IR 800 dyes. Thus a single gel can be used to score multiple markers. Scoring fragments is easy when patterns are simple and/or when segregation data is available that allows allele assignment. However for multi-allelic patterns, scoring can be more difficult. In these cases it is important to keep in mind that our assays are for phenotypes that we are trying to interpret as genotype. The banding pattern in gel (B) may be best scored as a phenotype or fingerprint <discuss>.

Scoring will also depend on the intended analysis. Gel (A) has been scored, above, using the “Mapmaker” notation where A = homozygous parent 1; B = homozygous parent 2; H = heterozygous; C and D are reserved for “not” a parent. In the case of lane 8 it is difficult to interpret whether the individual is an “A” or “H”. It is clear that it is not B, therefore we score it as C. Alternatively the gel can be scored in a two row/single column format where each row indicates an allele.
Genotyping Technology: Assays For SNPs

Schematic of the GoldenGate Assay for SNP genotyping as implemented using the Illumina BeadXpress.

Schematic of the ASPE assay as implemented using the Luminex system

Scoring Data for Luminex

The Luminex and BeadXpress collect data as mean fluoresence or pixels. These are then interpreted as alleles. A data file is provided as <COR96_005_(6-11-07).xls>. There are four different worksheets in the file. The first one, MFI, includes the raw data. For allele calling, if an intensity signal is less than 70 it is considered background, not a true signal; greater than 70 but less than 100 it is not
very strong signal, but I can accept it for allele calling; greater than 100 is considered a quality signal. The second worksheet contains bead count data. Bead count should be more than 100 and be similar between two beads of a marker; The third worksheet contains % data for allele calling calculated by the following formula:

\[
\frac{(\text{intensity signal of allele 1 or allele 2})}{(\text{total intensity signal from both allele 1 and allele 2})} \times 100
\]

The last worksheet shows allele calling. The allele calling is based on % data. Default cut off is 75%. For example, if allele 1 has 76% of total intensity signal and allele 2 has 24%, allele 1 will be called. If allele 1 has 75% and allele 2 has 25%, it is hetero. One must be cautious for this automatic allele calling (is the first case is really different from the second case?). So, quality control measures need to be put in place. For example if the data are based on a survey of inbred varieties, and a marker is frequently scored as a heterozygote data should be inspected and manually corrected. Likewise, segregating populations should be inspected for expected segregation ratios using a Chi-square test.

The **FlexMAP microsphere-based "Allele Specific Primer Extension" (ASPE) assay.**

1. Contigs are PCR-amplified from genomic DNA with "Locus-Specific Primers" ("FwdLSP" and "RevLSP").
2. For each SNP, two probes are designed. Each is a perfect match to one of the expected SNP alleles, with the nucleotide matching the polymorphic base at the 3’ end.
   The 5’ end of each probe carries a unique 24-base "FlexMAP tag", which is reverse-complementary to "FlexMAP anti-tags" uniquely assigned to different sets of microspheres (see 3 below).
   For the probe with a matching 3’-terminal nucleotide, primer extension occurs. Biotinylated dCTPs are incorporated into the newly polymerised DNA strand, allowing to label extended probes selectively with a streptavidin-linked reporter dye.
3. Probes are uniquely annealed to microspheres through the specificity of the FlexMAP "tag" / "anti-tag" recognition. The Luminex detection system identifies each microsphere by its internal dye, and records the associated reporter dye intensity (in the example "result table" as "Mean Fluorescence Intensity").
4. The reporter dye intensity detected for a pair of SNPprobes is used to calculate an "allele ratio" and interpreted as a genotype. In the illustrated example, the SNP's genotype is considered homozygous for allele "G".

**External Links**

Next Generation Sequencing:: [http://www.nature.com/nmeth/journal/v5/n1/full/nmeth1156.html](http://www.nature.com/nmeth/journal/v5/n1/full/nmeth1156.html)
Massively Parallel Sequencing:: [http://www.nature.com/nature/journal/v437/n7057/full/437326a.html](http://www.nature.com/nature/journal/v437/n7057/full/437326a.html)
GenomeTV: [http://www.youtube.com/user/GenomeTV](http://www.youtube.com/user/GenomeTV)
Next Generation Sequencing:  [http://www.youtube.com/watch?v=g0vGrNjpyA8&feature=related](http://www.youtube.com/watch?v=g0vGrNjpyA8&feature=related)
Center for Comparative Genomics, Western Australia   

**References:**