BIO 337

Tuesday, Feb 18 2014
Fred Sanger
13 August 1918 – 19 November 2013

Nobel Prize in Chemistry, 1958 for protein sequencing (insulin)
Nobel Prize in Chemistry, 1980 for DNA sequencing
Dideoxy sequencing

(a) Dideoxy nucleotides cannot form a phosphodiester band with the next incoming dNTP.

(b) DNA polymerase I + 4 dNTPs + ddATP

DNA sequence of original strand

Acrylamide gel
Automated dye-terminator sequencing

4-fluorescently labelled dideoxy dye terminators
- ddATP
- ddGTP
- ddCTP
- ddTTP

Pool and load in a single well or capillary
- scan with laser + detector specific for each dye
- automated base calling
- very long reads (~ 1000 bases)/run
Functional genomics by sequencing

Used sequencing chemistry invented by Fred Sanger in 1977

In the last 3-5 years, radically new sequencing approaches have been invented and employed for functional genomics, termed

- Next-generation sequencing (NGS, 2\textsuperscript{nd}, 3\textsuperscript{rd} generation)
- Ultra high-throughput sequencing
- Single-molecule sequencing
- Deep sequencing
that's a very compelling application for us, " says Tallon. of consumables in a reasonable amount of time, a 30 benchmarks in terms of being able to generate if it delivers as promised. "If they can meet their in a 2-hour run. Accordingly, many centers are with a slated sequencing throughput of 100 Gb 'Proton I' chip enables sequencing of 10 Gb in tems—the chips needed for each run, however, includes the instrument as well as the accom system itself costs roughly $250,000, which Proton may lead some to reconsider. The to adopt Ion Torrent technology, but the Ion Illumina-oriented workflows have hesitated to with improvements to MiSeq that deliver even 500 bases (two 250-base reads from either end with where we were a year or even 6 months ago, " says Michael Schatz, a bioinformatician at the indes, many data quality issues are steadily study—rendering the comparison invalid. And sequen- to 1000 users can upgrade for just $50,000, compared a major upgrade to the HiSeq. Although pared a major upgrade to the HiSeq. Although

Table 1  Next gen sequencing developers

<table>
<thead>
<tr>
<th>Company</th>
<th>Technology overview</th>
<th>On market?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Genomics</td>
<td>Optical analysis of arrays of ‘DNA nanoballs’</td>
<td>Yes</td>
</tr>
<tr>
<td>Genapsys Redwood City, California</td>
<td>Electronic detection of thermal/pH changes accompanying nucleotide addition</td>
<td>No</td>
</tr>
<tr>
<td>Genia Technologies</td>
<td>Pairing biological nanopores with semiconductor detection</td>
<td>No</td>
</tr>
<tr>
<td>GnuBio</td>
<td>Microfluidic system analyzes DNA nanodroplets with fluorescent primers</td>
<td>Alpha testing</td>
</tr>
<tr>
<td>Illumina</td>
<td>Sequencing by synthesis with fluorescently labeled reversible terminators</td>
<td>Yes</td>
</tr>
<tr>
<td>Lasergen Houston</td>
<td>Sequencing by synthesis with fluorescently labeled reversible terminators</td>
<td>No</td>
</tr>
<tr>
<td>Life Technologies (Ion Torrent)</td>
<td>Semiconductor sensor arrays detect protons released by nucleotide addition</td>
<td>Yes</td>
</tr>
<tr>
<td>NabSys Providence, Rhode Island</td>
<td>Single-molecule analysis revealing genomic location of sequencing probes</td>
<td>No</td>
</tr>
<tr>
<td>Noblegen Biosciences</td>
<td>Optical detection of ‘expanded’ DNA templates passing through synthetic pores</td>
<td>No</td>
</tr>
<tr>
<td>Oxford Nanopore Technologies</td>
<td>Detects changes in current as DNA strands pass through protein nanopores</td>
<td>No</td>
</tr>
<tr>
<td>Pacific Biosciences</td>
<td>Uses ‘zero-mode waveguides’ to optically detect real-time nucleotide addition</td>
<td>Yes</td>
</tr>
<tr>
<td>Qiagen (Intelligent Bio- Systems)</td>
<td>Sequencing by synthesis with fluorescently labeled reversible terminators</td>
<td>No</td>
</tr>
<tr>
<td>Roche (454)</td>
<td>Pyrosequencing of template-laden beads prepared by emulsion PCR</td>
<td>Yes</td>
</tr>
<tr>
<td>Stratos Genomics Seattle</td>
<td>Optical sequencing of fluorescently labeled, synthetically expanded templates</td>
<td>No</td>
</tr>
</tbody>
</table>
Table 2 Next-generation DNA sequencing instruments

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Cost per base</th>
<th>Read length (bp)</th>
<th>Speed</th>
<th>Capital cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Genomics</td>
<td>Low</td>
<td>Short</td>
<td>3 months</td>
<td>None (service)</td>
</tr>
<tr>
<td>HiSeq 2000 (Illumina)</td>
<td>Low</td>
<td>Mid</td>
<td>8 days</td>
<td>+++</td>
</tr>
<tr>
<td>SOLiD 5500xl (Life Technologies)</td>
<td>Low</td>
<td>Short</td>
<td>8 days</td>
<td>+++</td>
</tr>
</tbody>
</table>

**Maximum read length**

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Cost per base</th>
<th>Read length (bp)</th>
<th>Speed</th>
<th>Capital cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 GS FLX+ (Roche)</td>
<td>High</td>
<td>Long</td>
<td>1 day</td>
<td>+++</td>
</tr>
<tr>
<td>RS (Pacific Biosciences)</td>
<td>High</td>
<td>Very long</td>
<td>&lt;1 day</td>
<td>++++++++</td>
</tr>
</tbody>
</table>

**Maximum speed, minimum capital cost and minimum footprint**

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Cost per base</th>
<th>Read length (bp)</th>
<th>Speed</th>
<th>Capital cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 GS Junior (Roche)</td>
<td>High</td>
<td>Mid</td>
<td>&lt;1 day</td>
<td>+</td>
</tr>
<tr>
<td>Ion Torrent PGM (Life Technologies)</td>
<td>Mid</td>
<td>Mid</td>
<td>&lt;1 day</td>
<td>+</td>
</tr>
<tr>
<td>MiSeq (Illumina)</td>
<td>Mid</td>
<td>Long</td>
<td>1 day</td>
<td>+</td>
</tr>
</tbody>
</table>

**Combined prioritization of speed and throughput**

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Cost per base</th>
<th>Read length (bp)</th>
<th>Speed</th>
<th>Capital cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ion Torrent Proton (Life Technologies)</td>
<td>Low</td>
<td>Mid</td>
<td>&lt;1 day</td>
<td>++</td>
</tr>
<tr>
<td>HiSeq 2500 (Illumina)</td>
<td>Low</td>
<td>Mid</td>
<td>2 days</td>
<td>++++++++</td>
</tr>
</tbody>
</table>

a‘Low’ is < $0.10 per megabase, ‘mid’ is in-between and ‘high’ is > $1 per megabase. b‘Short’ is < 200 bp, ‘mid’ is 200–400 bp, ‘long’ is > 400 bp and ‘very long’ is > 1,000 bp. cEach “+” corresponds to ~$100,000. We list only commercialized instruments that can be purchased and for which performance data are publically available (as opposed to a comprehensive list of companies developing next-generation sequencing technologies). The categorizations refer to the aspect of sequencing performance to which the technology and/or its implementation in a specific instrument are primarily geared. These estimates were made at the time of publication, and the pace at which the field is moving makes it likely that they will be quickly outdated.
Next Generation Sequencing: Illumina

(a) Illumina’s library-preparation workflow

DNA fragments

1. Blunting by fill-in and exonuclease

2. Phosphorylation

3. Addition of A-overhang

4. Ligation to adapters

(b) Processing steps:

- OH OH
- Template hybridization
- Initial extension
- Denaturation
- First denaturation
- First cycle annealing
- First cycle extension
- Second cycle denaturation
- Second cycle annealing
- Second cycle extension

(c) Reversible dideoxy dye terminator

- Incorporate
- Detect
- Debloc
- Cleave fluor

- OH free 3' end

- 5' DNA
- 3' Fluor

- Cleavage site

- n = 35 total
Next Generation Sequencing: Illumina

Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Denature the double stranded molecules
Next Generation Sequencing: Illumina

**First chemistry cycle: determine first base**
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

**Image of first chemistry cycle**
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

**Before initiating the next chemistry cycle**
The blocked 3' terminus and the fluorophore from each incorporated base are removed.

Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

Paired-end and mate paired libraries

Figure 2

Comparison between (a) paired-end and (b) mate-pair sequencing library-construction processes.

Certainty can be obtained from longer read lengths, and several next-generation sequencers have offered increases in read length over time and refinement of their signal-to-noise characteristics to allow this certainty. Another fundamental improvement has resulted from so-called paired-end sequencing, namely producing sequence data from both ends of each library fragment. Read pairs can be obtained by one of two mechanisms: (a) paired ends or (b) mate pairs (Figure 2).

In paired-end sequencing, a linear fragment with a length of less than 1 kb has adapter sequences at each end with different priming sites on each adapter. The sequencing instrument is designed to sequence from one adapter priming site by use of the stepwise sequencing described above; then,
Emulsion PCR for clonal amplification

Used in next-gen sequencing by Roche/454 and Life/Applied Biosystems platforms
Pyrosequencing

1–2 million template beads loaded into PTP wells

Flow of single dNTP type across PTP wells

Flowgram

Ion Torrent pH based sequencing

(a) Structure of the Ion Torrent Ion Chip used in pH-based sequencing. (b) pH sensing of nucleotide incorporation.

Because the Ion Torrent sequencer uses native nucleotides for the sequencing reaction, there are no sources of noise akin to those identified for Illumina sequencing due to fluorescence or blocking groups on the reactants. Rather, noise accumulates due to phasing wherein not all the fragments are extended by nucleotide incorporation at each step. This effect is especially pronounced at sites in the library fragments with multiple bases of the same identity (so-called homopolymers). Coincidentally, the error model of Ion Torrent sequencing is defined largely by insertion or deletion errors that are also most prevalent at homopolymers. Here, the effect is most pronounced as the length of the homopolymer increases mainly due to loss of quantitation and ultimately saturation of the pH detector. Substitution errors also occur, albeit at very low frequency, and may be due to carryover effects from the previous incorporation cycle. Overall, the error rate of this instrument on a per read basis averages approximately 1% (i.e., 1 in 100 bases).

During the 2 years since the Ion Torrent was introduced, the average read length obtained has increased from 100 to 200 bp, produced as single-end reads. Unlike reversible terminator sequencing, the use of native nucleotides and the different sequences present on each bead loaded in the chip mean that incorporation rates differ from one bead to the next by incorporation cycle and according to sequence. As a result, a wide range of read lengths are obtained from any sequencing run, and this range increases as the total number of incorporation cycles increases. Throughput has increased over time, from 10 Mb per run average to 1 Gb per run average, by coupling longer reads with higher well density on the Ion Chip, which allows more beads to be sequenced per run. Each run requires approximately 2 h to complete; an intermediate series of washes requires an additional hour before the instrument can perform another run. Reaction volume miniaturization and the mass production of the Ion Chip using standard semiconductor techniques make this...
Single-molecule sequencing: Pacific Biosystems

Phospholinked hexaphosphate nucleotides

Limit of detection zone

Fluorescence pulse

Single-molecule sequencing: Pacific Biosystems

Nanopores can be

- **biological**: formed by a pore-forming protein in a membrane such as a lipid bilayer
- **solid-state**: formed in synthetic materials such as silicon nitride or graphene
- **hybrid**: formed by a pore-forming protein set in synthetic material
Features of next-gen sequencing

- Short reads (35 bp – 400 bp)
- Millions of reads per run ($10^7 – 5 \times 10^8$)
- Higher error rate per basepair raw
- No cloning in *E. coli*
- Huge amounts of data per experiment (20 GB primary/2 TB raw)
- Large data storage and computational analysis requirements
NGS data

Counting
- RNA-seq
- ChIP-seq
- etc.

Variants
- Cancer genomes
- Genetic variation
- etc.

Assembly
- New genomes, transcriptomes
- etc.
Cataloging variation between individuals in a species

Characterizing differences between cells within an individual

Describing the underlying cellular mechanisms

- Genome state: DNA methylation, Histone modifications, DNA-protein binding, Looping
- RNA life cycle: Transcription, RNA folding, Splicing, Genome replication, mRNA degradation, mRNA translation
- Genome folding

Sequencing the genome of a species

- Gorilla
- Neandertal
- Human
- Chimpanzee

Table 3

<table>
<thead>
<tr>
<th>Genetic variations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-nucleotide polymorphisms (SNPs)</td>
<td>Genetic differences between individuals</td>
</tr>
<tr>
<td>Copy-number variations</td>
<td>Genetic differences between individuals</td>
</tr>
<tr>
<td>Copy-neutral rearrangements</td>
<td>Genetic differences between individuals</td>
</tr>
<tr>
<td>Whole-gene duplications</td>
<td>Genetic differences between individuals</td>
</tr>
</tbody>
</table>

Figure 1

- Immunogenomics
- Cancer genomics
- Metagenomics

Where we are headed: a road map of sequencing science. The earliest sequencing projects focused on creating 'reference' genomes for individual species of interest. With new sequencing technologies and the advent of the '10,000 genome project' and larger initiatives, it is increasingly realistic to sequence a complete genome (that is, a reference genome) for a single individual. This has improved the capacity and local demand for next-generation DNA sequencing; the cost of generating, acquiring and/or storing samples; the cost of constructing and indexing fragment libraries; the cost of building and maintaining infrastructure for sequencing; the costs associated with transient or persistent mismatches between the local regulations for working with human subjects. Fortunately, using sequencing to identify genetic variation between individuals of a species is considerably easier than sequencing a complete genome. Nevertheless, the genome assemblies based purely on shotgun reads continue to fall short of the assemblies that can be achieved by hierarchical, clone-based Sanger sequencing. For example, the current 12-Mb genome assembly of Neandertal sequencing only requires the mapping of reads to a reference assembly while allowing for differences owing to polymorphisms or sequencing errors.
Some applications of next-gen sequencing

- Genome sequencing and variant discovery
- de novo assembly of bacterial and other small genomes
- DNA-protein interactions
- Chromatin and epigenetics
- RNA expression levels (profiling)
- ncRNA/small RNA discovery and profiling
- Metagenomics
- Sequencing extinct species (museomics)
Table 1 Applications of next-generation DNA sequencing

<table>
<thead>
<tr>
<th>Method</th>
<th>Sequencing to determine:</th>
<th>Example reference</th>
<th>‘Subway route as defined in Figure 3’</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA-Seq</td>
<td>A genome sequence</td>
<td>57</td>
<td>Comparison, ‘anatomic’ (isolation by anatomic site), flow cytometry, DNA extraction, mechanical shearing, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>Targeted DNA-Seq</td>
<td>A subset of a genome (for example, an exome)</td>
<td>20</td>
<td>Comparison, cell culture, DNA extraction, mechanical shearing, adaptor ligation, PCR, hybridization capture, PCR and sequencing</td>
</tr>
<tr>
<td>Methyl-Seq</td>
<td>Sites of DNA methylation, genome-wide</td>
<td>34</td>
<td>Perturbation, genetic manipulation, cell culture, DNA extraction, mechanical shearing, adaptor ligation, bisulfite conversion, PCR and sequencing</td>
</tr>
<tr>
<td>Targeted methyl-Seq</td>
<td>DNA methylation in a subset of the genome</td>
<td>129</td>
<td>Comparison, cell culture, DNA extraction, bisulfite conversion, molecular inversion probe capture, circularization, PCR and sequencing</td>
</tr>
<tr>
<td>DNase-Seq, Sono-Seq and FAIRE-Seq</td>
<td>Active regulatory chromatin (that is, nucleosome-depleted)</td>
<td>113</td>
<td>Perturbation, cell culture, nucleus extraction, DNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>MAINE-Seq</td>
<td>Histone-bound DNA (nucleosome positioning)</td>
<td>130</td>
<td>Comparison, cell culture, MNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>Protein-DNA interactions (using chromatin immunoprecipitation)</td>
<td>131</td>
<td>Comparison, ‘anatomic’, cell culture, cross-linking, mechanical shearing, immunoprecipitation, DNA extraction, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>RIP-Seq, CLIP-Seq, HITS-CLIP</td>
<td>Protein-RNA interactions</td>
<td>46</td>
<td>Variation, cross-linking, ‘anatomic’, RNase digestion, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, PCR and sequencing</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>RNA (that is, the transcriptome)</td>
<td>39</td>
<td>Comparison, ‘anatomic’, RNA extraction, poly(A) selection, chemical fragmentation, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>FRT-Seq</td>
<td>Amplification-free, strand-specific transcriptome sequencing</td>
<td>119</td>
<td>Comparison, ‘anatomic’, RNA extraction, poly(A) selection, chemical fragmentation, adaptor ligation, reverse transcription and sequencing</td>
</tr>
<tr>
<td>NET-Seq</td>
<td>Nascent transcription</td>
<td>41</td>
<td>Perturbation, genetic manipulation, cell culture, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, circularization, PCR and sequencing</td>
</tr>
<tr>
<td>Hi-C</td>
<td>Three-dimensional genome structure</td>
<td>71</td>
<td>Comparison, cell culture, cross-linking, proximity ligation, mechanical shearing, affinity purification, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>Chia-PET</td>
<td>Long-range interactions mediated by a protein</td>
<td>73</td>
<td>Perturbation, cell culture, cross-linking, mechanical shearing, immunoprecipitation, proximity ligation, affinity purification, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>Ribo-Seq</td>
<td>Ribosome-protected mRNA fragments (that is, active translation)</td>
<td>48</td>
<td>Comparison, cell culture, RNase digestion, ribosome purification, RNA extraction, adaptor ligation, reverse transcription, RNA depletion, circularization, PCR and sequencing</td>
</tr>
<tr>
<td>TRAP</td>
<td>Genetically targeted purification of poly-somal miRNAs</td>
<td>132</td>
<td>Comparison, genetic manipulation, ‘anatomic’, cross-linking, affinity purification, RNA extraction, poly(A) selection, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing</td>
</tr>
<tr>
<td>PARS</td>
<td>Parallel analysis of RNA structure</td>
<td>42</td>
<td>Comparison, cell culture, RNA extraction, poly(A) selection, RNase digestion, chemical fragmentation, adaptor ligation, reverse transcription, PCR and sequencing</td>
</tr>
<tr>
<td>Synthetic saturation mutagenesis</td>
<td>Functional consequences of genetic variation</td>
<td>93</td>
<td>Variation, genetic manipulation, barcoding, RNA extraction, reverse transcription, PCR and sequencing</td>
</tr>
<tr>
<td>Immuno-Seq</td>
<td>The B-cell and T-cell repertoires</td>
<td>86</td>
<td>Perturbation, ‘anatomic’, DNA extraction, PCR and sequencing</td>
</tr>
<tr>
<td>Deep protein mutagenesis</td>
<td>Protein binding activity of synthetic peptide libraries or variants</td>
<td>95</td>
<td>Variation, genetic manipulation, phage display, in vitro competitive binding, DNA extraction, PCR and sequencing</td>
</tr>
<tr>
<td>PhiT-Seq</td>
<td>Relative fitness of cells containing disruptive insertions in diverse genes</td>
<td>92</td>
<td>Variation, genetic manipulation, cell culture, competitive growth, linear amplification, adaptor ligation, PCR and sequencing</td>
</tr>
</tbody>
</table>

FAIRE-seq, formaldehyde-assisted isolation of regulatory elements—sequencing, MAINE-Seq, MNase-assisted isolation of nucleosomes—sequencing, RIP-Seq, RNA-binding protein immunoprecipitation—sequencing, CLIP-Seq, cross-linking immunoprecipitation—sequencing, HITS-CLIP, high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation, FRT-Seq, on-flowcell reverse transcription—sequencing, NET-Seq, native elongating transcript sequencing, TRAP, translating ribosome affinity purification. PhiT-Seq, phenotypic interrogation via tag sequencing.
Gene expression profiling with RNA-seq

Select RNA fraction of interest (poly(A), ribo-minus and others)

Fragment and reverse transcribe

Sequence, map onto genome

Quantitate (relative, absolute, nonmolar and others)

3x  2x  1x
Finding copy number variants with NGS

Deletions and amplifications with paired end sequencing

1. **Genomic DNA**
   - Fragment, select 3 kb fragments, ligate biotinylated adaptors

2. **Ligate to promote circularization**

3. **Digest with restriction enzyme, select fragments containing biotin**

4. **Sequence, map ends to reference genome sequence**

   - **Reference genome sequence**
   - **Study genome**
     - **No variant**
     - **Deletion**
     - **Insertion**
Whole-genome sequencing to identify disease gene
Genotyping to confirm disease allele

A. SH3TC2 Genotype and Phenotype

B. Results of R954X Genotyping

(by Restriction Fragment Length Polymorphism – RFLP)
Chromatin immunoprecipitation (ChIP)-seq

Once the experiment-specific peak shift has been estimated, the forward and reverse profiles are shifted and summed to produce the combined density profile (CDP) on one for reverse tags. These profiles are characterized by areas of strong enrichment where tags are particularly dense (which all subsequent analyses are carried out on). By combining the profiles in this manner, QuEST accomplishes two key computational tasks: first, it identifies candidate TFBS locations; second, it evaluates false positives and negatives. A 'peak', a putative TFBS, is predicted by the algorithm if the density profile value at the coordinate exceeds a user-specified threshold.
overrepresented motifs are graphically depicted (Weblogos). For each of the three transcription factors, significantly motif number within 100 bp of the peak.

For example, 89% of CDP peaks that contained a matching canonical activator functions. Thus, a peak with a score of 50 is due to a TFBS that was twice as strong as a peak with a score of 25.

The high resolution of QuEST peak calls is noteworthy. For two factors, as reflected by its almost threefold larger number of regulators of basic cell biology, rather than specialized factors with specific physiological roles. GABP is the more broadly acting of the two factors, as reflected by its almost threefold larger number of significant enrichment of genes that are involved in basic cellular processes, particularly those related to gene expression. These results are consistent with both GABP and SRF being fundamental activators of gene expression.

Go analysis (data not shown). We applied the same strategy to the NRSF dataset, which reproduced the discovery of NRSF half-sites which is consistent with previous results.

The score QuEST generates for each peak, according to which the motif center and 56% were within 10 bp of the genome. The motif scores for each peak are modeled as a Poisson random variable with a mean equal to the concentration of fragments that bind to the motif multiplied by the local sequence conservation. This model accounts for many known factors, such as DNA sequence composition, which can affect the physical affinity between TFs and DNA. Additional factors, such as DNA-protein cross-reactivity with other proteins as determined by western blot analysis (data not shown).
in performance between the NRSF and GABP datasets came from the Kharchenko's spp package, wtd and mtc, which were less sensitive in the GABP dataset. The decreased sensitivity of the spp methods on the GABP dataset may be caused by the broader enrichment regions noted in this dataset (see Figures S6, S7 and S8 and further discussion in the 'Spatial Resolution' section).

Directional scoring methods are known to be less useful for identifying broad enrichment signals, such as histone modification or RNA polymerase binding, due to blurring of the signal between the forward and reverse reads (Figure 1B).

Though high in confidence, the qPCR gold-standards cover only a handful of sites across the genome, perhaps limiting our ability to assess more subtle differences in sensitivity. To gain a more comprehensive picture of sensitivity between these methods, a whole genome scan for the presence of high confidence canonical binding motifs was conducted. This approach, which permits an assessment of sensitivity from a larger database, generated a list of more than 3000 potential NRSF and 6500 GABP binding sites. The coverage of these motif occurrences largely recapitulates the patterns seen with the qPCR binding site analysis, suggesting that the similarities observed with the high confidence qPCR database are not simply artifacts of the small sample size (Figure 5B,D). In summary, the sensitivity of all methods on the NRSF dataset remains remarkably similar over most of the peak-lists, while more noticeable differences emerge in examining the GABP data. The similarities from the NRSF data likely emerge from the fact that many algorithms may have been tested and trained on this same dataset, thereby optimizing their default settings. The differences seen with GABP highlight the potential variability in performance and seem to indicate that, for this dataset, directional scoring methods were less sensitive (SISSRS, mtc, wtd), corroborating the findings from our qPCR analysis.

It is important, however, to consider that high confidence motif sites represent putative binding sites for the transcription factor. Some sites may not be occupied under the experimental conditions and may not even be present in the cell line's genome, given that cell lines are prone to genomic instability. Thus, while the co-occurrence of motif instances and detected peaks likely represent true binding sites, the failure to identify a peak at a motif site has several possible explanations.

Specificity. Assessing the rate of false positives in the peak lists is a challenging task. The available set of qPCR-determined negative sites for NRSF provides only 30 ''true negatives'', defined as sites where enrichment was less than 3 fold [45]. By this standard, nine of eleven programs called a total of two putative false positives (CisGenome and QuEST found none). The same two ''true negative'' sites (chr20: 61280784–61280805 and chr6:108602345–108602365 in hg18) were identified by all nine programs. Although this could indicate some systematic bias in peak calling, Kharchenko et al. argue that, based on sequence tag distributions, these sites are likely bound by NRSF under the ChIP-seq experimental conditions (see Supplementary Fig. 9 from Kharchenko et al. [31]). Thus, we find these ''negative'' sites and their corollaries in the GABP dataset unreliable for assessing the specificity of the different programs using metrics such as a receiver operator curve (ROC), despite the fact that other groups have used this metric previously [12].

In the absence of an appropriate dataset for rigorous false positive testing, many investigators prefer to examine a stringent set of binding sites. Thus, programs must provide accurate means for ranking peaks according to some confidence metric. To assess...
the different statistics provided by the same program can produce similar tightly clustered, showing relatively little variation with respect to motif content in ranked peaks for the other two datasets were comparable to other programs (Figure S3). This result suggests that their ranking algorithms performed better on noisier datasets.

To assess how different these peak lists were, those peaks identified by programs on our simulated datasets at increasing noise thresholds. We note that the absence of a strong motif occurrence does not necessarily mean that the peak is not a true binding site. In some cases, high confidence motif occurrences could represent true binding sites with weak or non-canonical binding motifs. Nonetheless, high confidence motif occurrences within peaks are a good indicator of an actual binding event and therefore provide insight into the performance of different ChIP-seq methods.

In summary, we demonstrate the significance of peak rank in recovering high confidence motif sites. Our results support the findings of previous studies, such as those of Laajala et al. [12], which simply reports the average significance of motif searches of peaks.
Copy number variation has also been described as a mechanism to map in eukaryotic genomes (Snyder et al., 2010). Each has its own strengths and weaknesses, but typically all four are used to help identify SVs. SVs are elements that can be found in genomes that can be mapped using high-throughput sequencing (Korbel et al., 2007). However, these techniques offered 3D mapping of DNA interactions, which helped to understand the structural organization of genomes.

Figure 2

Sequencing technologies and their uses. Various NGS methods can precisely map and quantify chromatin features, DNA modifications and several specific steps in the cascade of information from transcription to translation. These technologies can be applied in a variety of medically relevant settings, including uncovering regulatory mechanisms and expression profiles that distinguish normal and cancer cells, and identifying disease biomarkers, particularly regulatory variants that fall outside of protein-coding regions. Together, these methods can be used for integrated personal omics profiling to map all regulatory and functional elements in an individual. Using this basal profile, dynamics of the various components can be studied in the context of disease, infection, treatment options, and so on. Such studies will be the cornerstone of personalized and predictive medicine.

interacting sites
Chromosome-DNA methylation
RRBS Bisulfite treatment creates C to U modification that is a
interaction
protein–DNA machinery and
Transcriptional
Genomic Sequencing). Identifying damaging polymorphisms
ibility markers and inherited disease traits (see ‘Medical
lowered data turn-around time associated with NGS have
individuals have gathered pace, and whole-genome sequen-
corganisms, projects to characterize the DNA sequence of
In addition to the sequencing of the genomes of different
Genome sequence and structural variation
significant challenges (Snyder (Pagani
organisms have been sequenced. As of June 2012, according to
sufficiently commonplace that a large number of different
variables are structural variations (SVs): large (SNP) project (Abdulla
Project (Ball
and B). Early microarray experiments indicated that SVs were
sequence (Consortium, 2003; Frazer
sequence genomes
accuracy will greatly enhance our abilities to accurately
difficult and leads to short contigs. Increasing read length and
assembly can be attempted from short reads, but this remains
(insertions and deletions) and larger structural variants are
sequencing platforms have different biases and abilities to call
precise mapping of variants between individuals.

### Table I The various NGS assays employed in the ENCODE project to annotate the human genome

<table>
<thead>
<tr>
<th>Feature</th>
<th>Method</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcripts, small RNA and transcribed regions</td>
<td>RNA-seq</td>
<td>Isolate RNA followed by HT sequencing</td>
<td>(Waern et al, 2011)</td>
</tr>
<tr>
<td>CAGE</td>
<td></td>
<td>HT sequencing of 5'-methylated RNA</td>
<td>(Kodzius et al, 2006)</td>
</tr>
<tr>
<td>RNA-PET</td>
<td></td>
<td>CAGE combined with HT sequencing of poly-A tail</td>
<td>(Fullwood et al, 2009c)</td>
</tr>
<tr>
<td>ChIPRP-Seq</td>
<td></td>
<td>Antibody-based pull down of DNA bound to IncRNAs followed by HT sequencing</td>
<td>(Chu et al, 2011)</td>
</tr>
<tr>
<td>GRO-Seq</td>
<td></td>
<td>HT sequencing of bromouridinated RNA to identify transcriptionally engaged PolII and determine direction of transcription</td>
<td>(Core et al, 2008)</td>
</tr>
<tr>
<td>NET-seq</td>
<td></td>
<td>Deep sequencing of 3' ends of nascent transcripts associated with RNA polymerase, to monitor transcription at nucleotide resolution</td>
<td>(Churchman and Weissman, 2011)</td>
</tr>
<tr>
<td>Ribo-Seq</td>
<td></td>
<td>Quantification of ribosome-bound regions revealed uORFs and non-ATG codons</td>
<td>(Ingolia et al, 2009)</td>
</tr>
<tr>
<td>Transcriptional machinery and protein–DNA interactions</td>
<td>ChIP-seq</td>
<td>Antibody-based pull down of DNA bound to protein followed by HT sequencing</td>
<td>(Robertson et al, 2007)</td>
</tr>
<tr>
<td>DNAse footprinting</td>
<td></td>
<td>HT sequencing of regions protected from DNase1 by presence of proteins on the DNA</td>
<td>(Hesselberth et al, 2009)</td>
</tr>
<tr>
<td>DNAse-seq</td>
<td></td>
<td>HT sequencing of hypersensitive non-methylated regions cut by DNase1</td>
<td>(Crawford et al, 2006)</td>
</tr>
<tr>
<td>FAIRE</td>
<td></td>
<td>Open regions of chromatin that is sensitive to formaldehyde is isolated and sequenced</td>
<td>(Giresi et al, 2007)</td>
</tr>
<tr>
<td>Histone modification</td>
<td></td>
<td>ChIP-seq to identify various methylation marks</td>
<td>(Wang et al, 2009a)</td>
</tr>
<tr>
<td>DNA methylation</td>
<td>RRBS</td>
<td>Bisulfite treatment creates C to U modification that is a marker for methylation</td>
<td>(Smith et al, 2009)</td>
</tr>
<tr>
<td>Chromosome-interacting sites</td>
<td>5C</td>
<td>HT sequencing of ligated chromosomal regions</td>
<td>(Dostie et al, 2006)</td>
</tr>
<tr>
<td>ChIA-PET</td>
<td></td>
<td>Chromatin-IP of formaldehyde cross-linked chromosomal regions, followed by HT sequencing</td>
<td>(Fullwood et al, 2009a)</td>
</tr>
</tbody>
</table>
The HiSeq X™ Ten, composed of 10 HiSeq X Systems, is the first sequencing platform that breaks the $1000 barrier for a 30x human genome.
Raw NGS data (FASTQ file)

4 lines per sequence

<table>
<thead>
<tr>
<th>Read ID</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>@HWI-ST1097:104:D13TNACXX:4:1101:18100:2240 1:Y:0:CAACTA</td>
<td>TGAGGCAAAACCCAACTTTATATGGGTCATATATAATGGTAAAGAAGGTTTAAA</td>
</tr>
<tr>
<td>+ =7=&lt;+2&lt;AACAA&lt;A+&lt;A97AB7&lt;7+2?ABBA@@B4A1?7A&lt;*::;00=AAA</td>
<td></td>
</tr>
<tr>
<td>@HWI-ST1097:104:D13TNACXX:4:1101:18326:2181 1:N:0:CAACTA</td>
<td>CATACATCAATTTTTTACAAAACTCGAATCTCGGTGGTATTATTCCGACAG</td>
</tr>
<tr>
<td>+ CCCFFFFFFHHHHHGJJIIGJGHJIHGIGHGDEIGHHFFHIIGG&gt;?DH6</td>
<td></td>
</tr>
<tr>
<td>@HWI-ST1097:104:D13TNACXX:4:1101:18259:2224 1:N:0:CAACTA</td>
<td>CAGGTGGAGGGACCCTGAATCTCGGTGGTATTATTCCGACAG</td>
</tr>
<tr>
<td>+ @C@DBDFDHHGHHI&lt;DHGH1CFGHIIBHIIEHIIGFDGAGE&gt;GGHGC@E</td>
<td></td>
</tr>
<tr>
<td>@HWI-ST1097:104:D13TNACXX:4:1101:18256:2243 1:N:0:CAACTA</td>
<td>GATAGGTGGTGGTGATCTAATTTGGTGGAACCTGGTGATATTTATGTATGATA</td>
</tr>
<tr>
<td>+ @@@FFFDDDFDFHGIJGIJJICGIJGIJAHGGGHBGGHGGHGEADHG</td>
<td></td>
</tr>
<tr>
<td>@HWI-ST1097:104:D13TNACXX:4:1101:18728:2073 1:N:0:CAACTA</td>
<td>TTTCTTTCTGAGAACCCTCTTTCTCAGTTTTGTGCTACTCTGAGGCAA</td>
</tr>
<tr>
<td>+ @@DDDDDHHD1&lt;AEHGGGG&lt;FHGIHEHEH9CDDA*??D&lt;DDHHAG&lt;?1?1?</td>
<td></td>
</tr>
</tbody>
</table>

Filesize: few hundred Mb to 1 or 2 Gb, ~100 million lines for one experiment!
Encoding quality scores

Quality character !"#$%&'()*/+,.-0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ
|         |         |         |         |
| ASCII Value | 33 | 43 | 53 | 63 | 73 |
| Base Quality (Q) | 0 | 10 | 20 | 30 | 40 |
| (ASCII-33) |

Probability of error = $10^{-Q/10}$

This is a **Phred** score, a standard measure of sequencing quality

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>
## NGS aligners

<table>
<thead>
<tr>
<th>Mapper</th>
<th>O.S.</th>
<th>Seq. Plat.</th>
<th>Input</th>
<th>Output</th>
<th>Min. RL</th>
<th>Max. RL</th>
<th>Mis-matches</th>
<th>Indels</th>
<th>Gaps</th>
<th>Splicing</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFAST</td>
<td>Linux,Mac</td>
<td>I,So,4, Hel</td>
<td>(C)FAST(A/Q)</td>
<td>SAM TSV</td>
<td>*</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Blat</td>
<td>Linux,Mac</td>
<td>N</td>
<td>FASTA</td>
<td>TSV BLAST</td>
<td>115000K</td>
<td>Score</td>
<td>Score</td>
<td>Y</td>
<td></td>
<td>De novo</td>
</tr>
<tr>
<td>Bowtie</td>
<td>Linux,Mac,Windows</td>
<td>I,So,4, Sa,P</td>
<td>(C)FAST(A/Q)</td>
<td>SAM TSV</td>
<td>41K</td>
<td>Score</td>
<td>Score</td>
<td>N</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>Linux,Mac,Windows</td>
<td>I,4,ion</td>
<td>FASTA/Q</td>
<td>SAM TSV</td>
<td>45000K</td>
<td>Score</td>
<td>Score</td>
<td>Y</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>BS-Seeker2</td>
<td>Linux, Unix, Mac</td>
<td>I</td>
<td>FASTA/Q, qseq</td>
<td>SAM BAM</td>
<td>10</td>
<td>200</td>
<td>Score</td>
<td>Y</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>BWA</td>
<td>Linux,Mac,Windows</td>
<td>I,So,4, Sa,P</td>
<td>FASTA/Q</td>
<td>SAM</td>
<td>4</td>
<td>200</td>
<td>Y</td>
<td>8</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>CloudBurst</td>
<td>Linux,Mac,Windows</td>
<td>N</td>
<td>FASTA</td>
<td>TSV</td>
<td>1K</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>ELAND</td>
<td>Linux, Unix, Mac</td>
<td>I</td>
<td>FASTA</td>
<td>TSV</td>
<td>15</td>
<td>150</td>
<td>2</td>
<td>Score</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>GMAP</td>
<td>Linux,Unix,Mac,Windows</td>
<td>I,4,Sa,Hel,ion,P</td>
<td>FASTA/Q, Native</td>
<td>SAM GFF</td>
<td>8*</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td>De novo</td>
</tr>
<tr>
<td>MapReads</td>
<td>Linux,Mac,Windows</td>
<td>So</td>
<td>FASTA/Q</td>
<td>TSV</td>
<td>10</td>
<td>120</td>
<td>Score</td>
<td>0</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>MAQ</td>
<td>Linux,Mac</td>
<td>I,So</td>
<td>(C)FAST(A/Q)</td>
<td>TSV</td>
<td>8</td>
<td>63</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>MOSAIK</td>
<td>Linux,Unix,Mac,Windows</td>
<td>I,4,Sa,Hel,ion,P</td>
<td>(C)FAST(A/Q)</td>
<td>BAM</td>
<td>15</td>
<td>1000</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>MrFAST</td>
<td>Linux,Unix</td>
<td>I</td>
<td>FASTA/Q</td>
<td>SAM DIVET</td>
<td>25</td>
<td>1000</td>
<td>Score</td>
<td>4</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Novoalign(CS)</td>
<td>Linux, I,So,4,Hel,Ion</td>
<td>I</td>
<td>(C)FAST(A/Q)</td>
<td>Illumina SAM Native</td>
<td>1</td>
<td>250</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Lib</td>
</tr>
<tr>
<td>RMAP</td>
<td>Linux,Mac</td>
<td>I,So,4</td>
<td>(C)FAST(A/Q)</td>
<td>BED</td>
<td>1110K</td>
<td>Y</td>
<td>0</td>
<td>N</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>SHRiMP2</td>
<td>Linux, Unix, Mac</td>
<td>I,So,4</td>
<td>FASTA/Q</td>
<td>SAM</td>
<td>301K</td>
<td>Y</td>
<td>Score</td>
<td>N</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>SOAP2</td>
<td>Linux</td>
<td>I</td>
<td>FASTA/Q</td>
<td>SAM TSV</td>
<td>271K</td>
<td>2</td>
<td>0</td>
<td>Y</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>SOAPSplice</td>
<td>Linux,Mac</td>
<td>I,4</td>
<td>FASTA/Q</td>
<td>TSV</td>
<td>133K</td>
<td>5</td>
<td>2</td>
<td>Y</td>
<td></td>
<td>De novo</td>
</tr>
<tr>
<td>SSAHA2</td>
<td>Linux,Mac</td>
<td>I,4,Sa</td>
<td>FASTA/Q</td>
<td>SAM</td>
<td>1548K</td>
<td>Score</td>
<td>Score</td>
<td>N</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>TopHat 2</td>
<td>Linux,Mac</td>
<td>I</td>
<td>FASTA/Q</td>
<td>BAM</td>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td></td>
<td>De novo</td>
</tr>
<tr>
<td>VMATCH</td>
<td>Linux,Mac</td>
<td>N</td>
<td>FASTA</td>
<td>TSV</td>
<td></td>
<td></td>
<td>Score</td>
<td>Score</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Aligning to a reference genome

### a Spaced seeds

**Reference genome (> 3 gigabases)**

- Chr1
- Chr2
- Chr3
- Chr4

**Short read**

ACTCCGTACTCTAAT

**Extract seeds**

- Position N
- Position 2
- CTGC CGTA AACT AATG

**Position 1**

- ACTG **** AAAC ****
- **** CCGT **** TAAT
- ACTG **** **** TAAT
- **** **** AAAC TAAT
- ACTG CCGT **** ****
- **** CCGT AAAC ****

**Index seed pairs**

- Seed index (tens of gigabytes)

**Confirm hits by checking “****” positions**

### b Burrows-Wheeler

**Reference genome (> 3 gigabases)**

- Chr1
- Chr2
- Chr3
- Chr4

**Short read**

ACTCCGTACTCTAAT

**Concatenate into single string**

**Burrows-Wheeler transform and indexing**

**Bowtie index (~2 gigabytes)**

**Look up ‘suffixes’ of read**

**Hits identify positions in genome where read is found**

**Convert each hit back to genome location**

**Report alignment to user**

---

*Nature Biotechnology (2009) 27: 455-457*
Variant calling with GATK

NGS DATA PROCESSING

Raw Reads

Non-GATK

Mapping
Duplicate Marking
Local Realignment
Base Quality Recalibration
Analysis-ready Reads

VARIANT DISCOVERY AND GENOTYPING

Sample 1 Reads → ... → Sample N Reads

Call Variants

SNPs, Indels, Structural Variations

INTEGRATIVE ANALYSIS

Raw Variants, External Data

Variant Quality Recalibration
Genotype Refinement
Variant Evaluation
Analysis-ready Variants

www.broadinstitute.org/gatk/
Variant calling

Chr 1 | 159,066,500 | 159,067,000 | 159,067,500 | 159,068,000 | 159,068,500

CD244

A/G
RNA-seq alignments

Processed mRNA

Mapping to genome
RNA-Seq with reference

FASTQ Files

Read Alignment (Bowtie/Tophat)

Reference Genome FASTA

Genome Browser (IGV)

Transcript annotation GTF/GFF

Read pre-filtering (ea-utils, FASTX tools, khmer)

Quality and error filtered, non-redundant FASTA

De-novo transcriptome assembly (Oases/All-paths)

De-novo Transcript contigs

Alignment BAM

Calculate Transcript Counts (HTseq)

Raw Transcript Counts

Normalization/Bias correction

Normalized Transcript Counts (TPM/RPKM/FPKM, UQUA)

Differentially expressed gene lists

RNA-Seq without reference (de-novo transcriptome assembly)

Current Opinion in Chemical Biology (2013) 17:4-11
What lies ahead...?

- End-to-end genome sequencing
- Sequencing entire pedigrees
- Sequencing within intact cells
- Single-cell genomes, transcriptomes, epigenomes
- Protein-protein interactions by sequencing
- Cell fate mapping
- Single molecule protein sequencing