Article:
Sethi, N, MacLennan, K, Wood, HM et al. (1 more author) (2016) The past and future impact of next-generation sequencing in head and neck cancer. Head & Neck, 38 (S1). E2395-E2402. ISSN 1043-3074

https://doi.org/10.1002/hed.24085

Title Page

Title:

The past and future impact of next generation sequencing in head and neck cancer

Authors:

Mr Neeraj Sethi, MBChB [1]

Professor Kenneth MacLennan, PhD[1]

Dr Henry M. Wood, PhD[1]

Professor Pamela Rabbitts, PhD[1]

[1]Leeds Institute of Cancer & Pathology, Wellcome Trust Brenner Building, St James' University Hospital, Leeds, West Yorkshire, UK. LS9 7TF

Corresponding author:

Mr Neeraj Sethi

Leeds Institute of Cancer & Pathology, Wellcome Trust Brenner Building, St James' University Hospital, Leeds, West Yorkshire, UK. LS9 7TF

Email: neerajsethi@doctors.org.uk

Tel: 07980 – 281 223

Fax: 0113 – 259 8885

Keywords: Next generation sequencing, head and neck cancer, Sanger sequencing, human papilloma virus, genomics

Financial declaration:

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/hed.24085

NS is funded by Leeds Charitable Foundation, the Mason Medical Research Foundation and the Royal College of Physicians & Surgeons of Glasgow.

Competing interests declaration:

None

Running title:

Impact of next generation sequencing

Abstract

Progress in sequencing technology is intrinsically linked to progress in understand cancer genomics. This review aims to discuss the development from Sanger sequencing to next generation sequencing (NGS) technology. We highlight the technical considerations for understanding reports using NGS. We discuss the findings of studies in head and neck cancer using NGS as well as the Cancer Genome Atlas. Finally we discuss future routes for research utilising this methodology and the potential impact of this.

Introduction

Progress in cancer research has paralleled that of progress in the various technologies that can be utilised and exploited. One of the most remarkable developments in the last decade has been the advent of next generation sequencing (NGS) technology. The human genome sequence was published in 2001[1, 2]. This ushered in a new era of scientific research, in which the correlation between genomic and phenotypic characteristics of disease could be made in new and promising ways.

Fearon and Vogelstein demonstrated that morphological development of colorectal cancer occurs in parallel with a stepwise progressive accumulation of genetic alterations[3]. Califano *et al* created a similar model for the genetic basis of head and neck cancer[4]. Since these landmark papers, one of the key theories driving cancer research has been that studying genetic changes across the entire genome (genomics) to identify alterations responsible for carcinogenesis and metastasis, could lead to new therapies and insights into how to manage patients with cancer.

The aim of this review is to explain the technological advances in sequencing and review their impact and discoveries thus far in head and neck cancer as well as discuss potential for the future.

Sanger sequencing

Sanger et al described Sanger sequencing in 1977[5]. This involved the copying of a template strand of DNA into radiolabelled complementary DNA (cDNA) strands. The synthesis of these strands is randomly terminated, and the sequence reconstructed from the final base of each strand[5-7].

The first genome to be sequenced was that of the bacteriophage phi X 174 (ΦX174)[8]. This utilised Sanger technology to identify the 5386 nucleotides. Sanger sequencing is accurate but can only sequence DNA fragments up to 1,000 bp in length. This would need to be performed 3 million times in order sequence the human genome once. For limited sequencing, however it is very cheap.

A progression in the rate of sequencing was achieved with Shotgun Sanger sequencing. This utilised plasmid cloning to produce cDNA fragments for sequencing, allowing longer overall templates of DNA to be sequenced more rapidly. The first cellular organism genome to be published was Haemophilus influenza in 1995[9]. This utilised Shotgun sequencing to reveal the 1,830,140 base pairs. Shotgun sequencing was key in increasing the speed at which DNA could be sequenced, and was the workhorse approach that produced the first draft of the human genome[1].

In addition to the laborious techniques and short sequences, Sanger sequencing was also limited in the accuracy of the first 40 and last 100 bases to be sequenced due to primer binding. The accuracy is also affected by increasing levels of guanine-cytosine (GC) content in the DNA strands to be

sequenced. Similarly, repetitive regions of DNA could also affect accuracy of sequencing[10]. Though shotgun sequencing did enhance the rate of sequencing, and latterly this became more automated, it still suffered similar issues[6].

Next generation sequencing

This describes a technology that differs from Sanger sequencing and represents a huge step forward in terms of speed of sequencing. It is important to understand that next generation sequencing (NGS) does not automatically mean whole genome sequencing (WGS) or exome sequencing (see table 1). It is a technology as opposed to a specific application.

NGS involves the breaking up of a DNA sample into many millions of fragments of known average length (see figure 1). Synthetic DNA “adaptors” are then bound to these fragments and labelled with an index primer (these are then referred to as DNA “libraries”). These fragments are then bound to a support matrix where an amplification reaction takes place followed by cycles of sequencing, which occur in parallel (leading to the term massively parallel sequencing). Signals are detected according to the nucleotides sequenced. Each DNA strand sequenced is termed a read. NGS has the capacity to produce hundreds of millions of reads. These are generally short (50 – 200 bases) and the huge numbers of reads requires considerable specialised computer resources to align these to the reference genome. The number of times the same area the genome is sequenced is referred to as depth (or

coverage) of sequencing. To produce more reliable data the same area needs to be sequenced many times according to the type of information required. For instance, whole genome sequencing requires higher coverage compared to copy number variation sequencing. There are however different platforms with variations in their chemistry. Each of these platforms can be used for sequencing DNA or RNA (see table 2).

Roche

The Roche GS-FLX 454 Genome Sequencer was the first commercially available platform (2004). These use an emulsion of beads as the matrix to which DNA libraries are bound. The amplification process ensures approximately 1 million copies of the same DNA fragment are bound to each bead. Nucleotides are then added and cDNA stands are synthesised via a pyrophosphate reaction (therefore this is often referred to as pyrosequencing). This reaction produces a light signal proportional to the nucleotides detected by a camera and converted to sequencing "reads" by a computer[11]. The use of this to sequence an individual's complete genome was published in 2008[12]. This represented a huge drop in the cost of sequencing a person's genomes (less than $1 million compared to more than $100 million by Venter et al)[1, 12]. Compared to other NGS platforms the Roche 454 is fast (23 hrs) and produces long reads (up to 1000 bp), though it cannot produce as many reads (therefore Roche data is low in depth). There is also a benchtop version (the Roche GS Junior).

More importantly, Roche announced in 2013 that it plans to shut down production of the 454 platform, though it will continue supporting current 454 sequencers already in use until 2016.

Illumina

The Illumina HiSeq uses a specialised glass slide called a flow cell as the matrix to which adaptor-ligated DNA is bound[13]. These fragments are then amplified to form clusters of identical DNA fragments. Fluorescent-labelled nucleotides are added to allow sequencing-by-synthesis and the signal released measured by a camera and translated to sequencing reads. This platform produces much more data than the Roche 454 in terms of depth of reads, though this does require experienced bioinformatics support. On high-output mode it takes two weeks to run, though this can be modified. A cheaper, quicker model (MiSeq) can be used for targeted sequencing of a smaller region of the genome[14]. Illumina have also introduced both benchtop HiSeq version and a larger machine to widen options in terms of cost and throughput.

Life Technologies

The Supported Oligonucleotide Ligation and Detection (SOLiD) platform uses magnetic beads to bind DNA libraries and undergo amplification by PCR. Four fluorescently labelled probes are added and ligate to the DNA library strands in a cyclical manner producing a signal, which is read by a camera. This is very precise in reading bases (99.99% accuracy) and produces good depth of

reads, but can be relatively prone to errors due to technical issues preparing the libraries and running the system[11, 14].

Life Technologies also produce the Ion Personal Genome Sequencer (PGM). This utilises semiconductor technology (ion torrent technology) in a similar fashion to pyrosequencing. Known nucleotides are introduced and hydrogen ions are released if they are added to the cDNA strand. These produce a pH change, which is detected and proportional to the number of bases added[14, 15].

Single Molecule Sequencing

This involves sequencing single molecules of DNA without any amplification. The advantages of this are removal of any potential bias or inaccuracy produced by the amplification step, as well as potentially increased accuracy, speed of sequencing and reduced cost[14, 16, 17]. The Helicos Heliscope system is still based on sequencing-by-synthesis and fluorescence detection[16]. Oxford Nanopore Technologies is developing a system of single molecule sequencing utilising a lipid bilayer, porous membrane that DNA molecules adhere to and then pass through on application of an electric current[17]. The passage of different bases through a pore produces alterations in the current across the membrane, which is measurable[18].

This type of sequencing is sometimes referred to as third generation sequencing, and accuracy of these platforms is still under investigation.

Technical Considerations

In addition to the different platforms there are several technical considerations to understand in the production and analysis of NGS data.

Laboratory

The source of the nucleic acid is important. Cell lines, fresh tissue and formalin-fixed paraffin embedded (FFPE) tissue are all potential sources. Cell lines can enable replicable results though genomic differences between cell lines and primary human cells have been described[19]. Fresh tissue is a good source of high quality nucleic acid, though can be more time-consuming to obtain. The archives of FFPE tissue around the world present huge potential in terms of numbers of samples. They also offer the advantage that follow-up data is often more easily and rapidly available for these samples. This nucleic acid is degraded and can be more challenging to work with as well as containing artefacts from the formalin-fixation process[20]. Techniques have improved so that FFPE tissue is increasingly being used[21, 22].

The purity of the source cell type is important. Tumour samples frequently contain mixed populations of cancerous epithelial cells, normal epithelium, lymphocytes and stromal cells. These non-cancerous, non-epithelial cells also contain nucleic acid, which can create “noise” masking the signal of the target cell. Previously, a minimum of 70% tumour cell fraction had been thought of as necessary, though with NGS this issue can be tackled a number of ways. By increasing the depth of sequencing, anomalies that are only present in a

smaller fraction of the cells being sampled can be detected[23]. The HNSCC samples used by the Cancer Genome Atlas (TCGA) had a median tumour cell fraction of ~50% and it is likely that much lower fractions will still yield very useful information[24]. This issue can also be accounted for with the development of algorithms that can enable lower fraction genomic anomalies to be identified, even with lower numbers of reads[25].

Cost

Sequencing costs have dropped dramatically since 2001, as shown in figure 2. This data from the National Human Genome Research Institute (NHGRI) compares DNA sequencing costs to a hypothetical trend described by Moore's Law (this predicts the trend of doubling in computing power associated with a decrease in hardware costs)[26].

Bioinformatics

Though the costs of sequencing a sample of DNA have reduced considerably, the data produced requires varying amounts of analysis. This is a challenging, specialised skill. Both academic and commercial institutes, with an interest in NGS are currently investing heavily in bioinformaticians. This cost is often not accounted for in claims of the "$1,000 genome"[27].

Bioinformatics is key in the analysis of NGS data and in accounting for potential error. Sources of error in NGS include PCR artefact. Many NGS methodologies involve one or more PCR steps, during which errors in PCR replication can cause mismatches in the alignment to the reference genome,

causing essentially a false positive. Similarly PCR steps inevitably produce duplicates of the same segments of nucleic acid. These waste sequencing reads and if there are excessive amounts reduce the accuracy of sequencing overall[28, 29]. Inaccuracy in the sequencing platform calling (recognising) bases is also an issue. This is referred to as sequencing error and varies in reports from 1 in 1000 bases to 1 in 10,000,000 bases[30, 31]. Though these appear low, given the billions of bases sequenced with each run this is significant. Attempts to reduce this error include increasing read depth (the number of times each DNA strand is sequenced), using technical replicates (sequencing the same library repeatedly to identify error) and biological replicates (multiple samples from the same cell type to identify random errors and repetitive abnormalities)[28, 32].

The primary aim of the bioinformatician is to process and analyse the raw NGS data with accurate 'calling' of anomalies (whether mutation, copy number etc) and minimising the rate of false positives. The degree of variation for cancer genomes compared to the reference genome varies considerably. Adjustments must therefore be made for the sample's background anomaly rate, ploidy and purity[23]. For example if a sample contained 50% tumour DNA and a mutation is present on one arm of a triploid chromosome, this will only be present in 16.6% of the sequenced reads[23]. The depth of sequencing will influence the ability to detect a mutation such as this as will the presence of a matched normal sample, also sequenced at sufficient depth. An error can be made due to detecting a germline event in the tumour and failing to detect it in

the normal or when a mutation is mistakenly called in the tumour when both the tumour and normal are wild-type[23, 25].

The presence of important low frequency mutations in clonal subpopulations within the sample is another confounding issue. Sequencing depth and the use of algorithms that are stable in the presence of data from genomically heterogeneous tumours such as HNSCC is essential.

New methods of analysing NGS to produce more accurate results or to discover clinically relevant patterns are produced every month[33-35]. Much of this data is essentially open source and available for download e.g. CNAnorm, a programme available from Bioconductor.org designed to estimate copy number aberrations in cancer samples[36]. Considerable effort is required to keep abreast of these as well as the ongoing results of sequencing being published.

Specifically for head and neck cancer, the Mutant Allele Tumour Heterogeneity algorithm (MATH) was developed to measure intratumour heterogeneity from publically available exome sequencing data[37, 38]. A higher MATH measure was found to be associated with specific groups of head and neck cancer with poorer outcome (those with *TP53* mutations, HPV-negative and HPV-negative tumour with increased smoking pack-year history)[37].

NGS and head and neck cancer

The first major studies in the use of NGS in HNSCC were published in 2011[38, 39]. These two studies together performed whole exome sequencing on 106 patients with HNSCC in total. These included oral, oropharyngeal, laryngeal, hypopharyngeal and sinonasal tumours. It also included HPV-positive and negative tumours. These studies confirmed the findings of previous genomic work that *TP53* was the most commonly mutated gene in HNSCC and also discovered the second most commonly mutated gene was *NOTCH1* (in around 15% of patients)[38, 39]. This was the first time *NOTCH1* had been implicated in HNSCC.

Interestingly these studies also found that HPV-positive tumours had approximately half the mutation rate of HPV-negative tumours[38, 39]. On analysing subgroups they also found smokers had a higher rate of guanosine to thymidine point mutations, in addition to having a higher rate of mutations. In general they found around 130 mutated genes per sample. The surprisingly low proportion of recurring mutations could be related to the mix of subsites reducing the number in each group, but gives a picture that each head and neck tumour is genomically quite different to the next.

In a follow up publication by Lui *et al* in 2013 a further 45 tumours had undergone whole exome sequencing, making a total of 151 sequenced tumours available for analysis[40]. Again a large number of mutated genes were identified per sample and a high degree of inter-tumour mutational heterogeneity observed. Developing their analysis, they focused on specific functional pathways that had previously been identified as targetable in

cancer. By doing this they found 31% of HNSCC in their cohort contained phosphoinositide 3-kinase (*PI3K*) pathway mutations. This signalling axis has been shown to have a role in cancer cell growth, survival, motility and metabolism[41-43]. Lui *et al* found that *PI3K*-pathway mutated HNSCC contained a higher rate of mutations in known cancer genes and that those with concurrent mutations in *PI3K* pathway genes were all advanced tumours implicating his pathway in HNSCC progression[40]. This study highlighted the potential for NGS to identify therapeutic targets and biomarkers in HNSCC.

Integrative genomics is a burgeoning research area and the combination of NGS data with other techniques was demonstrated by Pickering et al who used exome sequencing in 40 OSCC patients with SNP array copy number data, gene expression and miRNA expression as well as DNA methylation. They identified four major driver pathways in OSCC including mitogenic signalling, Notch, cell cycle and *TP53*. Though a small group they also highlighted two subgroups defined by the key genes *FAT1* and *CASP8*[23]. This approach also identified currently and potentially targetable genomic anomalies.

The TCGA has performed comprehensive genomic analysis of 279 untreated HNSCC cases[44]. This included whole exome sequencing, whole genome sequencing and whole transcriptome sequencing as well as miRNA, DNA methylation and copy number profiling. Thirty-six of the tumours were HPV-positive and 243 were HPV-negative. The majority of tumours were oral cavity and laryngeal (n = 244/279, 87%). Of 33 oropharyngeal tumours they found

64% were HPV-positive, whilst only 6% of non-oropharyngeal tumours were HPV-positive[44].

The TCGA found HPV-positive and negative tumours to have an overall different mutation profile, with HPV-positive tumours exhibiting infrequent mutations in *TP53, CDKN2A, FAT1* and *AJUBA.* They found 86% of HPV-negative tumours harboured *TP53* mutations whilst only 1 of 36 HPV-positive tumours had a *TP53* mutation. Whilst PIK3CA was found to be mutated in both HPV-positive and negative tumours, a specific mutation of the helical domain of *PIK3CA* was predominant in HPV-positive tumours – an important finding when considering targetable events. *EGFR* was found to be rarely mutated in HPV-positive tumours compared to HPV-negative tumours[44]. This could have serious implications regarding the use of EGFR-inhibitors in these patients.

The larger numbers involved in the TCGA do lend a greater credence to their ability to analyse subgroups. They confirmed previously reported gene expression subtypes (atypical, mesenchymal, basal and classical)[44-46]. Using an integrated approach they were able to identify genomic markers and suggest pathways associated with each subtype.

The India Project Team of the International Cancer Genome Consortium (ICGC) demonstrated the advantages of concentrating resources and collaborative efforts by reported whole exome sequencing on 50 gingivo-buccal SCC (GBSCC) and targeted resequencing on a further 60 GBSCC[24]. It

is vital that genomic patterns identified in different cohorts of HNSCC are not mistakenly assumed to be present in another. The prevalence of betel quid chewing in South-East Asia means a different profile of HNSCC is seen in this region. This study identified 5 new genes associated with GBSCC and 3 molecular subgroups demonstrating different disease-free survival.

Increasingly, important therapeutic subgroups of patients with HNSCC will be discovered as the numbers of tumours being sequenced grows. This is important in the effort towards “personalised medicine”. Part of the revolution being driven by NGS will be the shift away from purely classifying tumours by pathologic criteria and integrating genomic subgroups that are clinically relevant and will guide treatment decisions. Gross et al took advantage of the TCGA data available (WES, copy number variation, mRNA and miRNA expression) and combined 250 HPV16 negative cancers, aged under 85[25]. They were able to link loss of 3p with TP53 mutation as a marker for significantly decreased survival (1.9 yr compared to >5 yr for TP53 mutation alone). They also identified mir-548k expression as an additional marker for further reduced survival.

Another study performed whole exome sequencing on 16 younger non-smokers with oral tongue cancer (<45 years old) and 28 older smokers[48]. Surprisingly, this study found the two groups to be genomically similar. On interrogating TCGA data for lung adenocarcinoma, bladder urothelial carcinoma and HNSCC, a smoking mutation signature was generated. Both young and older oral tongue cancers were found to be most similar to a non-

smoking mutation profile. Admittedly this is a small group of uncommon cancers but the combination of individual study data with TCGA data is a good example of the accumulative power of NGS.

Targeted sequencing could also be useful in confirming a cell line mutational profile when attempting to demonstrate in vitro efficacy of targeted therapies, though if course the lack of epigenetic factors must be borne in mind[49].

NGS also has applications for the determination of HPV-status. This technology can be used to detect copies of HPV DNA within the sample being sequenced. It also has the advantage that all sub-types of HPV can be screened for simultaneously[50]. This can be achieved with low-coverage and relatively low-cost NGS technology and can be performed as an additional analysis of the same sequencing data being obtained for other purposes at no extra cost. Issues with the use of this technology relate to the fact that detection of a single copy of HPV DNA within the sample does not mean the tumour was driven by HPV and there is no accepted standard for the number of detectable copies that should be regarded as a positive result. Work in cervical cancer certainly shows promise for a NGS based high risk HPV genotyping assay[51]. Conway et al found NGS to be comparable to PCR and *p16* immunohistochemistry with excellent specificity[50]. It has also been used to screen a large number of oral verrucous carcinoma samples for all subtypes of HPV establishing the scarcity of HPV in this type of oral cancer[52].

RNAseq has also been used to evaluation HPV16 expression in seven young patients (average age 37) with oral tongue tumours[53]. This study found that these patients had a poor prognosis and found no evidence of HPV16 expression. Seiwert et al compered targeted exome sequencing and copy number profiles of 51 HPV16 positive and 69 HPV16 negative tumours. They found a similar overall mutational burden in both groups though unique mutations in *DDX3X* and *FGFR2/3* were found in HPV16 positive tumours[44].

Parfenov et al used NGS to investigate the tumour-host interaction in HPV16 positive HNSCC[45]. They examined whole genome sequencing and DNA methylation profiles in 35 HPV positive tumours and compared these to 270 HPV16 negative samples from the TCGA cohort. Whole genome sequencing allowed them to identify sites of integration of HPV DNA into the host genome. By doing this they were able to identify cancer genes at the sites of integration that were potentially disrupted and involved in the carcinogenic mechanism in virally driven HNSCC.

The issue of intra-tumour heterogeneity has gained increasing prominence recently with landmark studies in renal cell carcinoma using NGS to demonstrate clearly significant mutational difference in different samples from the same tumour[54, 55]. The potential impact of this on the use of genomic biomarkers to guide treatment and clinical trials is huge. Three samples from a single oropharyngeal tumour and two samples from its corresponding cervical metastasis underwent whole genome sequencing in a study by Zhang et al[56]. This found only 41% of all somatic point mutations were shared across

all five samples. Though this concurred with larger studies clearly the high cost and singular workload in applying this technology is demonstrated with only the ability to analyse one tumour. This cost is continuing to come down but CNVseq or targeted sequencing of a smaller panel of known genes could be used to demonstrate genomic heterogeneity at lower cost.

Conclusion

NGS technology has revealed significant genomic characteristics of HNSCC. The technology available is advancing continually as are the methods for analysing the data produced. In light of this, it is important for raw NGS data obtained by different groups to be made publicly available after publication. The ability to add to the pool of data is vital for tumours that are less common such as HNSCC. The issue of subsite signatures and subgroups according to ethnicity, inheritance, HPV and smoking amongst others is also a reason to try and pool data in order to increase the power of available data. Projects such as Head and Neck 5000 present a fantastic opportunity for large numbers of tissue and blood to be interrogated, though these attempts need to be carefully planned to avoid wasting resources[57].

Precancer in HNSCC still requires analysis using NGS technology with comparison to spatially and temporally-related cancer in order to help divine tumour promoters and drivers.

Since the first draft of the human genome was produced the cost of whole genome sequencing has dropped from approximately a billion dollars to a couple of thousand dollars. The speed at which this data can be obtained has gone from years to two weeks. Advancements will continue to be made to improve accuracy and data processing. The information gleaned from NGS will be collated and combined with clinicopathologic data on an increasingly large scale. Combining NGS with other genomic approaches on a large scale will reveal biomarkers and therapeutic targets. This will enable the development of clinically–relevant, molecular sub-groups that will guide treatment.

Acknowledgements

NS is funded by Leeds Teaching Hospitals Charitable Foundation, the Mason Medical Research Foundation and the Royal College of Physicians & Surgeons of Glasgow.

We would like to acknowledge Dr Neil Hayes for his kind input regarding the TCGA HNSCC group work.

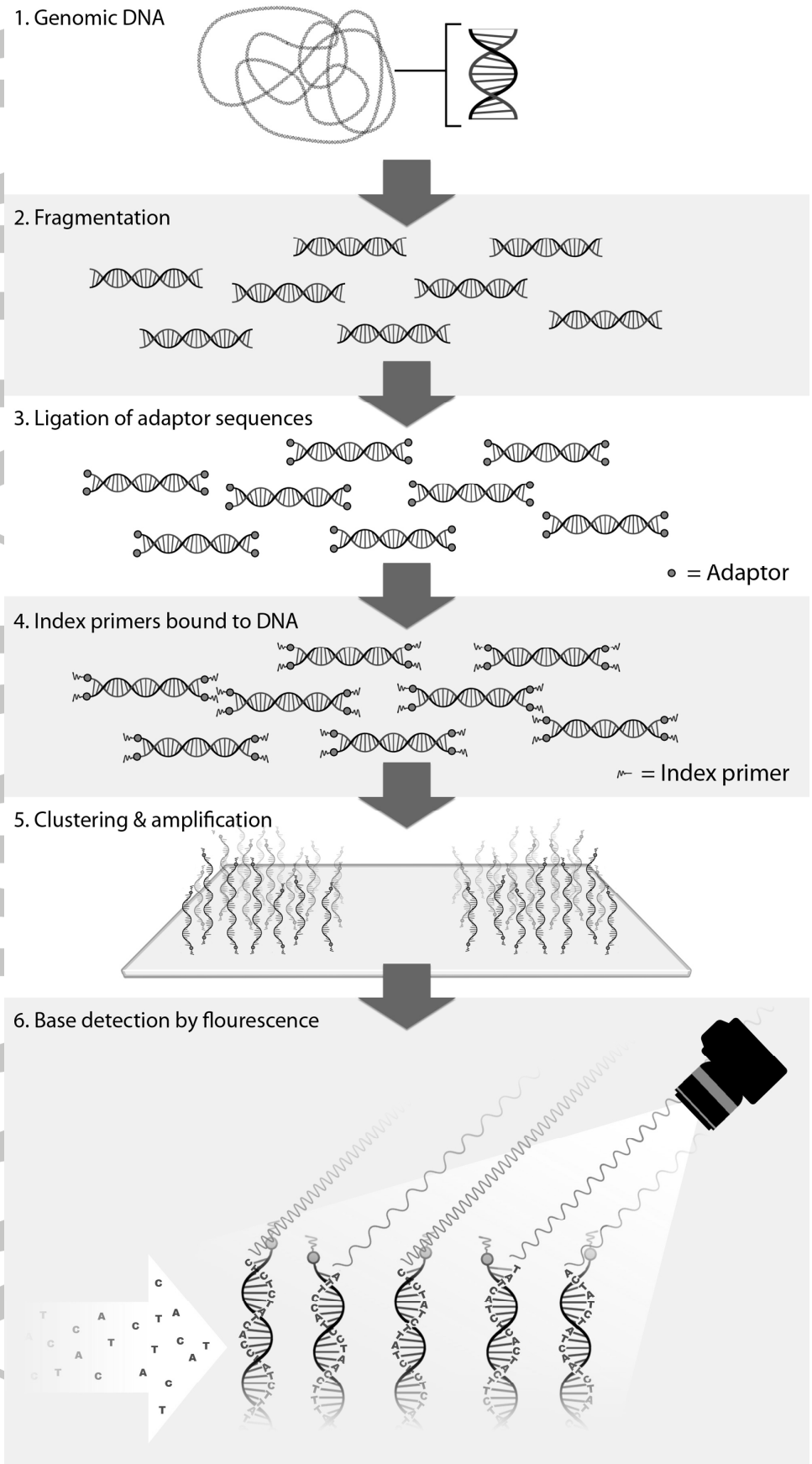
References

1. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science*. 2001; 291: 1304-51.
2. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409: 860-921.
3. Fearon ER and Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell*. 1990; 61: 759-67.
4. Califano J, van der Riet P, Westra W, et al. Genetic progression model for head and neck cancer: implications for field cancerization. *Cancer research*. 1996; 56: 2488-92.
5. Sanger F, Nicklen S and Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*. 1977; 74: 5463-7.
6. Franca LT, Carrilho E and Kist TB. A review of DNA sequencing techniques. *Quarterly reviews of biophysics*. 2002; 35: 169-200.
7. Swerdlow H and Gesteland R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic acids research*. 1990; 18: 1415-9.
8. Sanger F, Air GM, Barrell BG, et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*. 1977; 265: 687-95.
9. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*. 1995; 269: 496-512.
10. Bankier AT, Beck S, Bohni R, et al. The DNA sequence of the human cytomegalovirus genome. *DNA sequence : the journal of DNA sequencing and mapping*. 1991; 2: 1-12.
11. Zhang J, Chiodini R, Badr A and Zhang G. The impact of next-generation sequencing on genomics. *Journal of genetics and genomics = Yi chuan xue bao*. 2011; 38: 95-109.

12. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*. 2008; 452: 872-6.
13. Holt RA and Jones SJ. The new paradigm of flow cell sequencing. *Genome research*. 2008; 18: 839-46.
14. Morey M, Fernandez-Marmiesse A, Castineiras D, Fraga JM, Couce ML and Cocho JA. A glimpse into past, present, and future DNA sequencing. *Molecular genetics and metabolism*. 2013; 110: 3-24.
15. Rothberg JM, Hinz W, Rearick TM, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011; 475: 348-52.
16. Mardis ER. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*. 2008; 9: 387-402.
17. Branton D, Deamer DW, Marziali A, et al. The potential and challenges of nanopore sequencing. *Nature biotechnology*. 2008; 26: 1146-53.
18. Guy AT, Piggot TJ and Khalid S. Single-stranded DNA within nanopores: conformational dynamics and implications for sequencing; a molecular dynamics simulation study. *Biophysical journal*. 2012; 103: 1028-36.
19. Wilkening S, Stahl F and Bader A. Comparison of primary human hepatocytes and hepatoma cell line Hepg2 with regard to their biotransformation properties. *Drug metabolism and disposition: the biological fate of chemicals*. 2003; 31: 1035-42.
20. Wong SQ, Li J, Tan AY, et al. Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing. *BMC medical genomics*. 2014; 7: 23.
21. Wood HM, Belvedere O, Conway C, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic acids research*. 2010; 38: e151.
22. Schweiger MR, Kerick M, Timmermann B, et al. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. *PloS one*. 2009; 4: e5548.
23. Meyerson M, Gabriel S and Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nature reviews Genetics*. 2010; 11: 685-96.
24. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nature genetics*. 2013; 45: 1134-40.
25. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*. 2013; 31: 213-9.
26. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).
27. Mardis ER. Anticipating the 1,000 dollar genome. *Genome biology*. 2006; 7: 112.
28. Robasky K, Lewis NE and Church GM. The role of replicates for error mitigation in next-generation sequencing. *Nature reviews Genetics*. 2014; 15: 56-62.
29. Lou DI, Hussmann JA, McBee RM, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proceedings of the National Academy of Sciences of the United States of America*. 2013; 110: 19872-7.

30. Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S and Schuster SC. Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PloS one*. 2013; 8: e55089.
31. Peters BA, Kermani BG, Sparks AB, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*. 2012; 487: 190-5.
32. Ajay SS, Parker SC, Abaan HO, Fajardo KV and Margulies EH. Accurate and comprehensive sequencing of personal genomes. *Genome research*. 2011; 21: 1498-505.
33. Youn A and Simon R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*. 2011; 27: 175-81.
34. Hajirasouliha I, Mahmoody A and Raphael BJ. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*. 2014; 30: i78-i86.
35. Sun Z, Evans J, Bhagwate A, et al. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. *BMC genomics*. 2014; 15: 423.
36. Gusnanto A, Wood HM, Pawitan Y, Rabbitts P and Berri S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*. 2012; 28: 40-7.
37. Mroz EA and Rocco JW. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral oncology*. 2013; 49: 211-5.
38. Stransky N, Egloff AM, Tward AD, et al. The mutational landscape of head and neck squamous cell carcinoma. *Science*. 2011; 333: 1157-60.
39. Agrawal N, Frederick MJ, Pickering CR, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science*. 2011; 333: 1154-7.
40. Lui VW, Hedberg ML, Li H, et al. Frequent mutation of the PI3K pathway in head and neck cancer defines predictive biomarkers. *Cancer discovery*. 2013; 3: 761-9.
41. Samuels Y, Wang Z, Bardelli A, et al. High frequency of mutations of the PIK3CA gene in human cancers. *Science*. 2004; 304: 554.
42. Engelman JA, Luo J and Cantley LC. The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. *Nature reviews Genetics*. 2006; 7: 606-19.
43. Courtney KD, Corcoran RB and Engelman JA. The PI3K pathway as drug target in human cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2010; 28: 1075-83.
44. Network TCGA. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. 2015; 517: 576-82.
45. Chung CH, Parker JS, Karaca G, et al. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer cell*. 2004; 5: 489-500.
46. Walter V, Yin X, Wilkerson MD, et al. Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. *PloS one*. 2013; 8: e56823.

47. Lechner M, Frampton GM, Fenton T, et al. Targeted next-generation sequencing of head and neck squamous cell carcinoma identifies novel genetic alterations in HPV+ and HPV- tumors. *Genome medicine*. 2013; 5: 49.
48. Pickering CR, Zhang J, Neskey DM, et al. Squamous Cell Carcinoma of the Oral Tongue in Young Non-Smokers Is Genomically Similar to Tumors in Older Smokers. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2014.
49. Nichols AC, Yoo J, Palma DA, et al. Frequent mutations in TP53 and CDKN2A found by next-generation sequencing of head and neck cancer cell lines. *Archives of otolaryngology--head & neck surgery*. 2012; 138: 732-9.
50. Conway C, Chalkley R, High A, et al. Next-generation sequencing for simultaneous determination of human papillomavirus load, subtype, and associated genomic copy number changes in tumors. *The Journal of molecular diagnostics : JMD*. 2012; 14: 104-11.
51. Yi X, Zou J, Xu J, et al. Development and validation of a new HPV genotyping assay based on next-generation sequencing. *American journal of clinical pathology*. 2014; 141: 796-804.
52. Samman M, Wood H, Conway C, et al. Next-generation sequencing analysis for detecting human papillomavirus in oral verrucous carcinoma. *Oral surgery, oral medicine, oral pathology and oral radiology*. 2014; 118: 117-25 e1.
53. Bragelmann J, Dagogo-Jack I, El Dinali M, et al. Oral cavity tumors in younger patients show a poor prognosis and do not contain viral RNA. *Oral oncology*. 2013; 49: 525-33.
54. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England journal of medicine*. 2012; 366: 883-92.
55. Gerlinger M, Horswell S, Larkin J, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature genetics*. 2014; 46: 225-33.
56. Zhang XC, Xu C, Mitchell RM, et al. Tumor evolution and intratumor heterogeneity of an oropharyngeal squamous cell carcinoma revealed by whole-genome sequencing. *Neoplasia*. 2013; 15: 1371-8.
57. http://www.headandneck5000.org.uk.



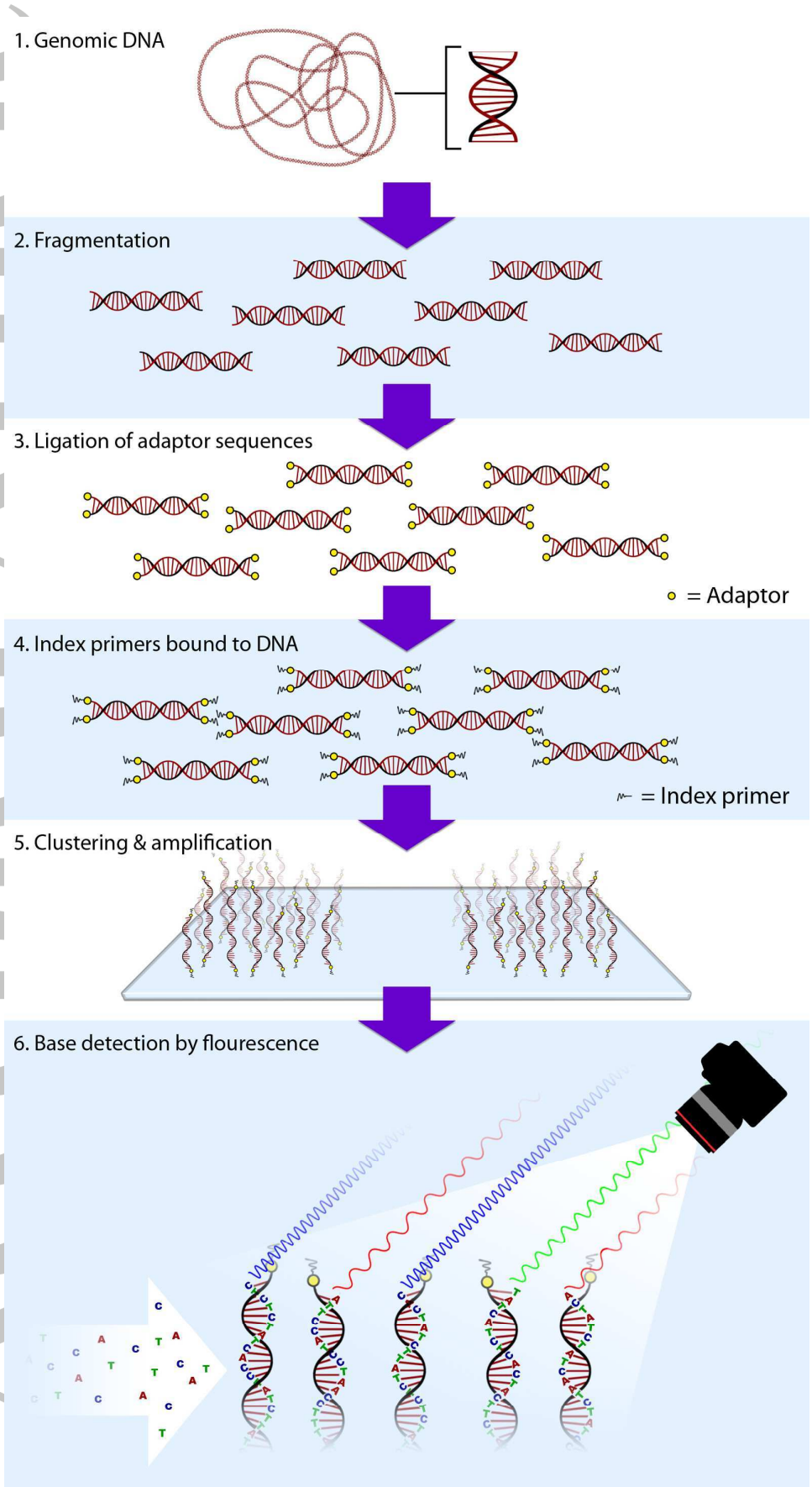
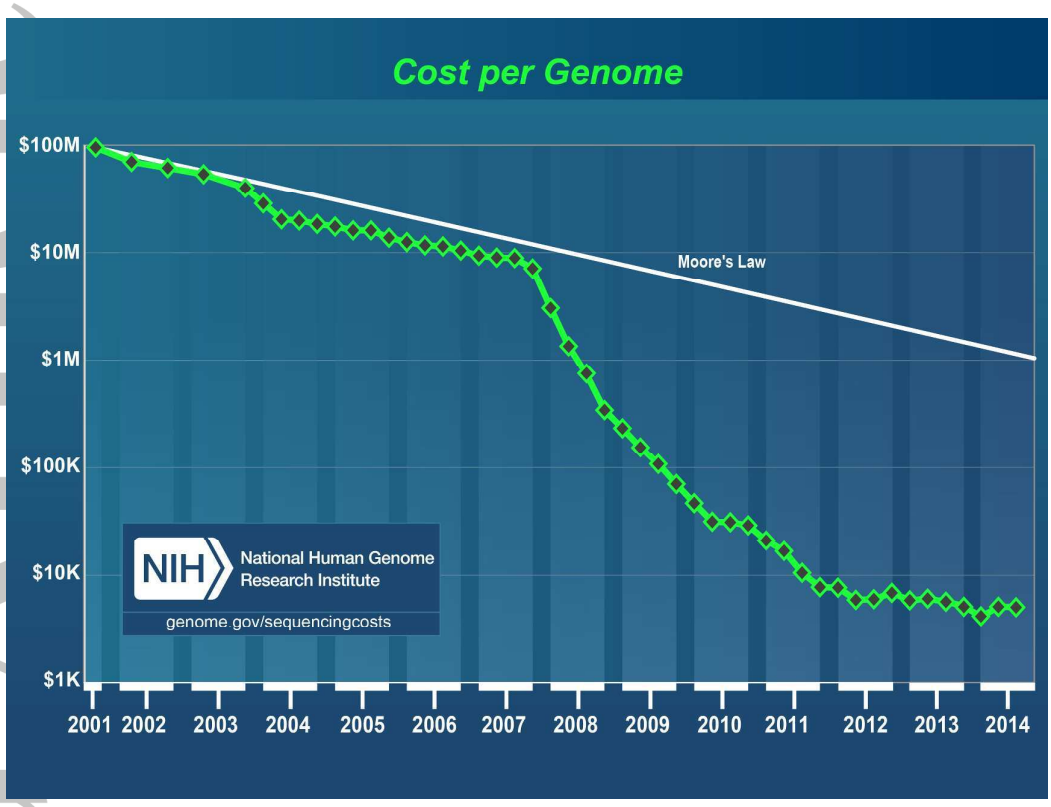


Figure 1: Above is shown the processing nucleic acid into a form for next generation sequencing. (1) The genomic DNA has been extracted from the tissue sample. (2) This is then broken down into fragements of approximately equal maximum length. This is necessary as NGS produces sequencing reads of a fixed maximum length, dependent on the platform and settings. (3) + (4) Adaptor sequences and primers are ligated to the fragmented DNA in order for this to bind to the sequencing matrix and for each strand to be identifiable when analysing the reads in the subsequent data. (5) The labelled DNA binds to a sequencing matrix and each strand undergoes an amplification process producing clusters which are all read many times, thus improving the accuracy of the sequencing. In the Illumina platform the matrix takes the form of a glass slide as shown above though this can take the form of bead, as in the Roche platform. (6) Nucleotides are added and cDNA strands are synthesised from these. A laser is used to make the nucleotides fluoresce. This signal is detected and converted into sequencing reads.



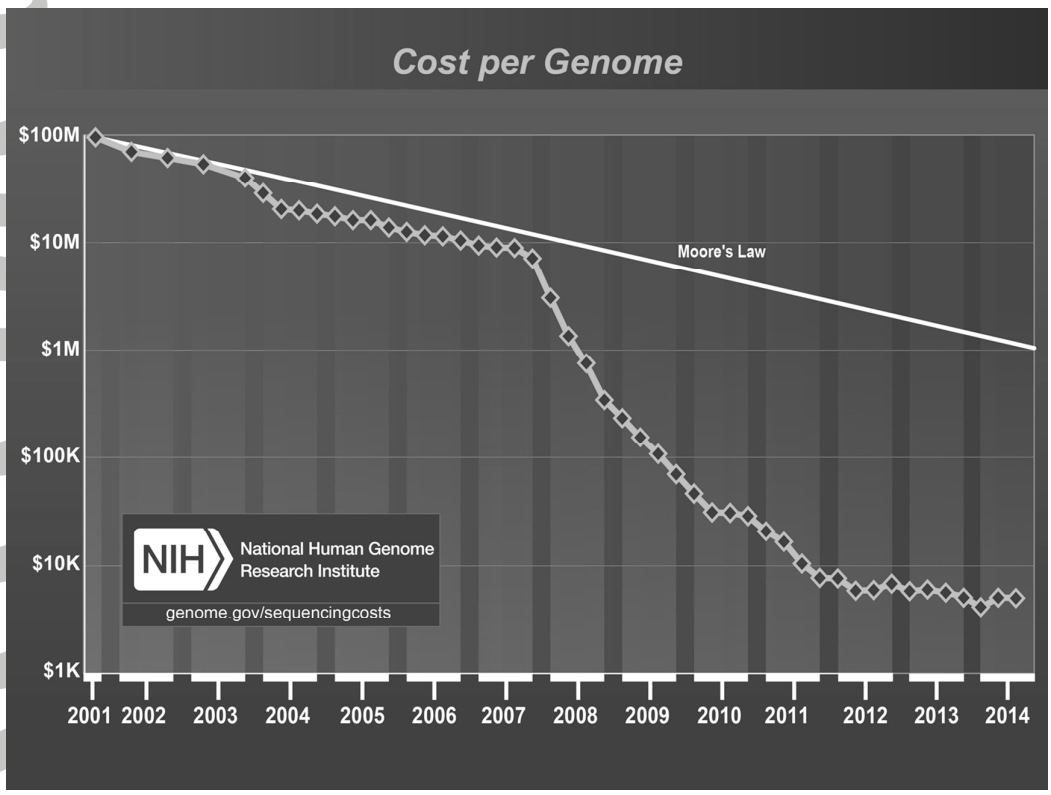


Figure 2: Illustration of the reduction in DNA sequencing costs (dotted line) since 2001 compared to a hypothetical trend dictated by Moore's Law (solid line). Reproduced with kind permission from KA Wetterstrand, DNA

Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed 18/12/14.

Term	Source Material	Information obtained
Whole Genome Sequencing	**DNA**	Entire DNA sequence
Whole Exome Sequencing	**DNA**	Sequences of all known exons for known genes
Whole Transcriptome Sequencing (RNASeq)	**RNA**	Sequence of all RNA molecules contained in the sample
Targeted sequencing	**DNA/RNA**	Sequence of a subset of genes of identified region of the genome
microRNA Sequencing (miRNASeq)	**microRNA**	Sequences of all known miRNAs
Copy Number Variation Sequencing (CNVSeq)	**DNA**	Areas of gain or loss in copy number of the genome

Table 1: Different type of sequencing performed with NGS technology

Platform	Underlying mechanism	Read length (bp)	Data output/run	Time/run	Advantages	Disadvantages
Roche 454	Pyrosequencing	700	0.7 Gb	24 hr	Fast Longer read length	High reagent cost Higher error rate in repetitive regions
Illumina HiSeq	Sequencing by synthesis	36 - 100	600 Gb	27 hr - 10 days	Higher data yield/run, Higher throughput, Cost effective in terms of data yield	Short read length Longer run time
SOLiD (Life Technologies)	Sequencing by ligation	35 – 75	180 Gb	7 day to 2 weeks	High accuracy Very high throughput	Long run time Short read length Complex sample preparation
Ion Torrent (Life Technologies)	Sequencing by synthesis	200	1 Gb	2 – 4 hr	Fast Low cost	Lower data yield High error rate in repetitive regions Short read length

Table 2: Comparison of currently available NGS platforms