Four study design principles for genetic investigations using next generation sequencing

Clinton C Mason

Division of Pediatric Hematology and Oncology, Department of Pediatrics, University of Utah, 417 Wakara Way, Salt Lake City, UT 84108, USA

Correspondence to: C Mason
clint.mason@hsc.utah.edu

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2017;359:j4069
http://dx.doi.org/10.1136/bmj.j4069

Accepted: 7 August 2017

Proper study design is crucial for obtaining meaningful results and reaching correct conclusions in scientific investigations. How to apply key study design principles to next generation sequencing (NGS) can be unclear and is sometimes overlooked by researchers who are new to this technology. A proper study design in NGS studies can be achieved with awareness of factors that can influence sequencing results and by taking the necessary steps to limit their biasing assessment. These steps include using simultaneously or similarly sequenced controls, randomising, determining the necessary sequencing depth, and performing a power calculation to determine the necessary sample size for testing hypotheses at significance levels adjusted for multiple comparisons. Applying study design principles to NGS studies will increase the likelihood of identifying genetic factors associated with human traits and diseases.

Studies using next generation sequencing (NGS) dominate genetic research, contributing to rapid increases in our understanding of nearly all diseases.[1] These studies are highly attractive for investigating the genetic features of almost any trait or malady owing to affordability and breadth of genomic assessment. Although NGS provides a wealth of information, it is not free from biases that can result in incorrect or limited conclusions. Yet potential obstacles can be overcome by correctly applying study design principles.

SUMMARY POINTS

- Next generation sequencing (NGS) enables extensive genetic assessment but is prone to artifacts and requires a proper study design
- Comparison with simultaneously or similarly sequenced controls can reduce artifacts
- Randomisation prevents a rise in false positives when the NGS process is changed (knowingly or unknowingly) during the study
- Power to assess the hypothesis in question depends on both the sample size and sequencing depth and should be calculated beforehand to determine the appropriate level of sample multiplexing

The first steps of planning an NGS study are identifying the hypothesis, the type of study that will test it most efficiently, and the appropriate technology to use. The goal is to limit biases and their resultant false positives, while having adequate power to identify true positive effects. NGS studies can be used to assess DNA variants and mutations, RNA transcript abundance, methylation levels, transcription factor binding, and knockdown or knockout effects through shRNA or CRISPR screens. The advantages and nuances of these applications have been reviewed thoroughly elsewhere,[1-4] but experimental design strategies applied to NGS have received less attention.[5-8] This paper seeks to help clinician investigators apply four key study design principles—similar assessment of controls, randomisation, sufficient evaluation, and adequate sample size—in conducting an NGS experiment, focusing on assessing DNA variants and mutations using targeted sequencing, whole exome sequencing (often abbreviated as WES), and whole genome sequencing (abbreviated as WGS, see box 1 for a glossary of terms) in case-control and cohort studies.

Similar assessment of controls

Determining the genetic variations that are associated with affected cases requires comparison with unaffected controls. Publicly available control databases are sometimes used in NGS studies as replacements for simultaneous controls to save costs. Although DNA mutations are more reproducible than other genetic features, such as expression or methylation,[9-12] simultaneously sequencing DNA from appropriate controls is still very useful, particularly in whole exome sequencing and other targeted sequencing studies.

In contrast to whole genome sequencing, which probes the entire genomic sequence nearly uniformly, whole exome sequencing and other targeted sequencing methods use commercially produced "bait libraries" to enrich certain portions of the genome for focused interrogation (such as all exons or certain genes). As bait libraries, reagents, and sequencing machines are routinely updated by manufacturers to enhance coverage, simultaneous controls are necessary for eliminating biases stemming from use of controls assessed with different variations of these components. Historical or database controls, such as the 1000 Genomes Project,[13] the NHLBI GO Exome Sequencing Project,[14] or the Exome Aggregation Consortium,[15] are likely to have been prepared with different reagents, sequenced at different depths, targeted to different regions, and processed with

Box 1: Glossary

- ***Artifact***—An undesired factor or bias preventing or limiting assessment of a hypothesis
- ***Control database***—A collection of genetic results from generally healthy or unselected participants in a previous study
- ***Depth of coverage***—The number of times a position in a genetic sequence has been assessed; aka sequencing depth, read depth
- ***DNA fragment***—A small portion (typically hundreds to thousands of consecutive bases) of DNA required for NGS assessment
- ***Germline variant***—A variation in the genetic sequence of an individual from that in the general population and present in the DNA of nearly all cells of the body due to its having been inherited or arising as an early de novo mutation in the individual
- ***Multiplexing***—A technique for assessing the genetic sequence of multiple samples simultaneously with reduced cost but also reduced depth of coverage
- ***Next generation sequencing (NGS)***—A technique for identifying genetic sequences by interrogating a large number of genetic fragments in parallel, often providing many assessments of the genetic sequence
- ***Read***—A typically small portion of a genetic sequence determined by a next generation sequencing machine "reading" (identifying) some or all of the bases from a single genetic fragment
- ***Read depth***—The number of times a position in a genetic sequence has been assessed; aka depth of coverage, sequencing depth
- ***Sequencing depth***—The number of times a position in a genetic sequence has been assessed; aka depth of coverage, read depth
- ***Somatic mutation***—A spontaneous change in the DNA sequence of any somatic cell that may proliferate and lead to cancer or other disease
- ***Study design***—Planning a research investigation that will allow meaningful statistical assessment of the hypothesis free from artifacts and biases
- ***Targeted sequencing***—An NGS method focused on identifying the genetic sequence at only specified regions (targets) of a DNA sample
- ***Whole exome sequencing (WES)***—An NGS method focused on identifying the genetic sequence in only the exonic (protein coding) regions of a DNA sample
- ***Whole genome sequencing (WGS)***—An NGS method for identifying the entire genetic sequence of a DNA sample

different bioinformatics software pipelines. Moreover, their ethnic and genetic make-up may differ from the samples under investigation. Without comparable controls, investigators may mistakenly assume that variants they identify in cases that have not previously been observed in database controls owing to differences in assessment are associated with disease, leading to false positives (fig 1). Statements to the effect of "the observed variant was not present in the 1000 Genomes database" thus provide limited information without knowing how well the location of that variant was sequenced in the 1000 Genomes project.

This does not preclude the use of control databases or historical in-house controls—these resources are extremely valuable for excluding many common variants or artifacts, particularly as running large cohorts of controls with every new investigation is impractical. Further, recent in-house samples that have been run on the same machines, baits, pipelines, and populations may be sufficiently similar to be used instead of strict "simultaneously run" controls.

Another exception is a two step study design where only affected cases are initially assessed using NGS and compared with historical or database controls to filter out many of the common variants or artifacts. Then the remaining unfiltered variants undergo secondary assessment in cases and new controls selected by the investigator simultaneously, using a cheaper sequencing method. This design can be more cost effective, particularly when the variant's rarity must be established in a large number of controls or when seeking to establish that a common variant endows a significant relative risk. Including some simultaneous controls in the first step may be less costly overall when false positives are sufficiently reduced before the second stage.

When somatic mutations are being sought in DNA from cancer tissue, additional simultaneous sequencing of DNA from normal tissue or other source of germline DNA from the same patient is the most valuable control. This enables common and rare germline variants (that will be identified in both the individual's tumour and normal tissue) to be distinguished from somatic mutations present only in the tumour DNA.

Example A

Situation—A clinician investigator wants to use whole genome sequencing to assess whether any DNA variants are associated with increased risk of onset of a particular rare disease, as well as identify the prevalence of translocations. Research cases are patients identified by referral with no known relationship to each other.

Application—The investigator should identify people without the disease as controls, ideally matched for age, geographic location, disease related exposures, ethnicity, and gender. DNA from both cases and controls should be sequenced at the same time. The investigator should filter out common variants detected in database controls as well as common variants identified in the simultaneous controls.

If not applied—Without simultaneous controls, the investigator cannot distinguish potential rare variants in the cases from new artifacts or common variants in previously undersequenced regions.

Example B

Situation—A clinician investigator studying a cancer cohort wants to determine the association between survival and mutations in known cancer related genes, using a targeted sequencing panel that focuses on nearly all suspect genes.

Application—In addition to extracting DNA from each patient's tumour, the investigator should also extract DNA from either the patient's healthy tissue or a suitable germline surrogate. Both the tumour and normal DNA should be sequenced at the same time and processed together in all subsequent analyses.

If not applied—Failure to run paired samples may result in artifacts and rare, private variants being mislabeled as somatic mutations in the tumour sample.

Randomisation

Randomisation prevents systematic differences in the experimental process from causing spurious genetic associations. For example, several factors can affect the output of NGS—including the sequencing machine

doi: 10.1136/bmj.j4069 | *BMJ* 2017;359:j4069 | the bmj

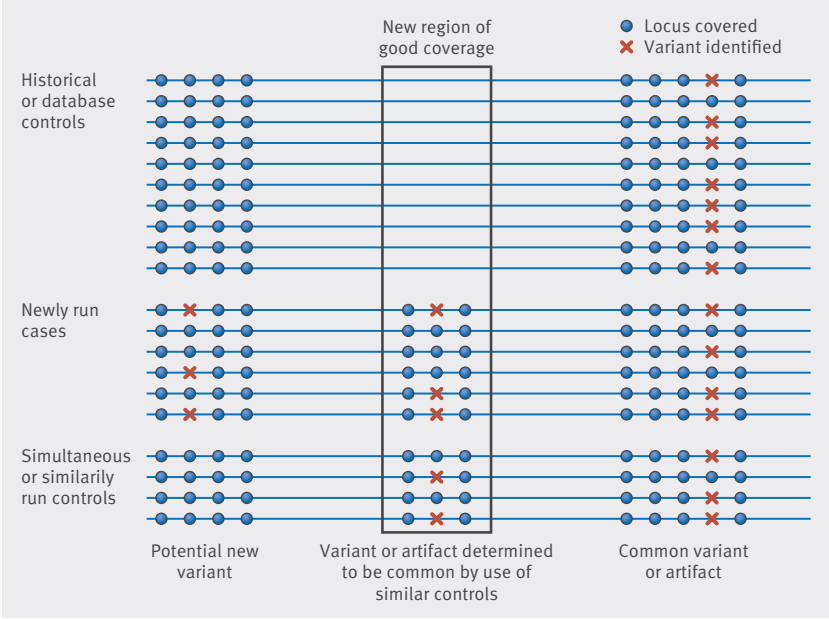


Fig 1 | Simultaneous or similar sequencing of controls is necessary in NGS studies. Locations in the human genome are illustrated horizontally, with each horizontal line representing an individual sample. Variants identified in new cases but not in historical or new controls might be disease related variants requiring further follow-up (left). Variants may not be identified in historical controls if the locus was not sequenced well previously, detection in newly run controls can prevent misidentification of common variants or artifacts as being potentially disease related (middle). When similarly good assessment occurs in all three cohorts, a common variant is identified in each (right).

having changes in clustering density over time and changes in reagents—potentially allowing biases to influence the results of non-randomised studies. Longer term studies are at risk of changes in bait libraries, software, and sequencing machines and of DNA degradation.

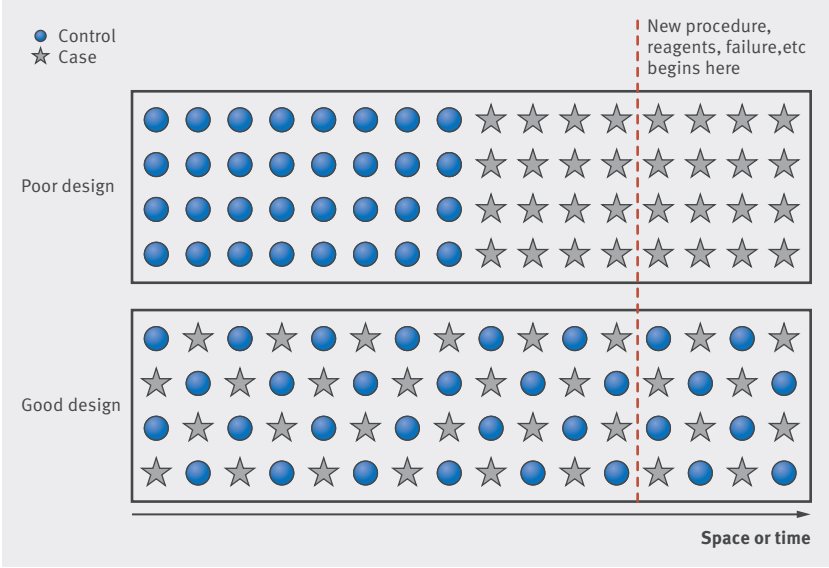


Fig 2 | Randomisation is necessary to avoid confounding or weakening of results by temporal or spatial changes in the NGS process. If the case and control samples are not randomised, temporal changes may disproportionately affect the groups. Randomisation lessens the possibility of such changes (whether known or unknown) biasing the sequencing results.

Spatial differences might also unduly affect non-randomised experiments. These can occur when different sequencing lanes or machines have systematically different total read yields due to differences in optics, clustering, or reagent flow. Fig 2 shows how cases and controls might be affected differently in non-randomised studies; a change in sequencing efficiency could disproportionately affect either cases or controls. But when the samples are randomised, the proportion of cases and controls affected by such a process change will be similar—reducing statistical power but not creating false positives.

Different randomisation strategies can be used.[16] Simple randomisation rearranges the samples without assessing whether the numbers of cases and controls is equal across potential confounding sources; for example, using a random number generator for ordering does not always cause cases and controls to be equally distributed. Block randomisation reduces potential confounding by requiring equal (or specified) numbers of cases and controls in each block, then simple randomisation is used within each block. Appropriate randomisation will reduce false positives and improve reproducibility.

Example C

Situation—A clinician investigator identifies a multigenerational family affected by a rare disease that reflects a Mendelian inheritance pattern. The disease phenotype presents in childhood, enabling accurate determination of affected status in adults. The investigator wants to use whole genome sequencing to sequence both affected and unaffected family members to identify potentially causal, inherited variants.

Application—After judiciously selecting which family members to sequence—for example, choosing affected members most distantly related to reduce the overall shared genome—the investigator obtains blood samples and places them in a randomised order. DNA extraction, sequencing, and analyses are performed on the samples in this order.

If not applied—If the investigator had sequenced all affected individuals first and unaffected family members much later, systematic changes to the process might have caused otherwise avoidable false positives.

Sufficient sequencing depth and multiplexing

NGS relies inherently on multiple assessments of each nucleotide. In whole exome and whole genome sequencing, DNA fragments from many cells are isolated, sequenced, aligned, and mapped to the genome. The sequencing depth (also referred to as read depth or coverage) is the number of times that these fragments provide information on the nucleotide base at a particular position—for example a locus might be sequenced by 15 reads or with 15× coverage. This depth can vary widely over a region, particularly for targeted sequencing methods.

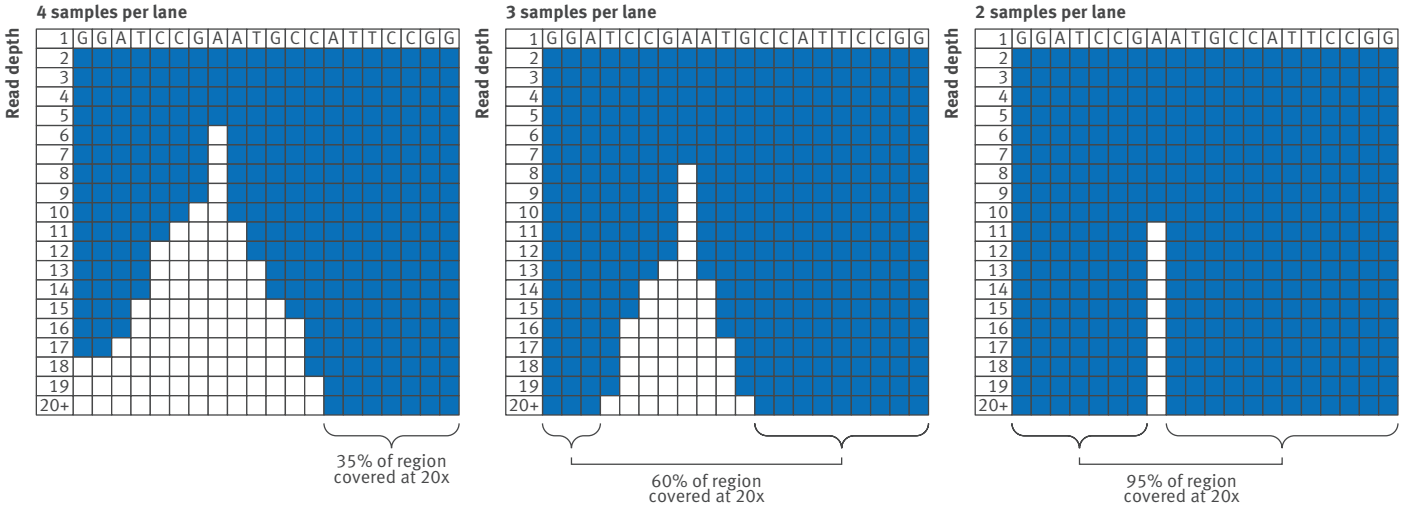
Some analytical methods allow evaluation and comparison of bases that have few reads, even when systematic depth differences exist between cases and

controls,[17] but many analysis pipelines completely filter out loci with low coverage, preventing the sample from contributing to hypothesis evaluation at bases with a read depth below a specified threshold. Thus, a substantial portion of the desired genomic region may not be determined in samples with overall low sequencing depth. The portion that is determined is often referred to as the percent of the target region covered or the percent covered at a particular depth. This percentage often conveys more practical information than the mean or median sequencing depth across a region.

Fig 3 (top panel) shows the total reads at each base across a genomic region of interest for the same sample run on a sequencing lane (multiplexed) with one, two, or three other samples. Multiplexing reduces the cost of sequencing a sample as a trade-off for reduced sequencing depth.[18] Multiplexing can occasionally cause misidentification of reads,[19] but it is generally considered useful because of its financial benefit. Investigators must decide the multiplexing for an NGS experiment that reflects a balance between the total number of samples that can be assessed for a fixed cost and the proportion of the target region that will be adequately sequenced. Optimal multiplexing can often be estimated from previous, similar experiments (fig 3 (bottom panel)).

For germline studies, 20 to 30 high quality reads are often deemed sufficient to confidently identify the presence or absence of a variant.[20] Several factors will influence the observed variant allele frequency in somatic studies, including purity of tumour or normal tissue, copy number variation, and extent of clonal development. Hence, confident detection of somatic mutations often requires much greater depth. Somatic whole exome sequencing discovery studies performed on many of the current lane based sequencers commonly multiplex no more than two samples in each lane, with goals of achieving 40-120× coverage over substantial portions of the exome. Targeted sequencing of on up to hundreds of specific genes or regions allows much deeper sequencing and much greater multiplexing owing to the reduced size of the target region. It can achieve depths of >1000×, enabling the detection of mutations present at low frequencies or the coverage of difficult regions.

Whole exome sequencing was recently found to be more cost effective than whole genome sequencing at detecting exonic, germline variants.[21] As sequencing costs decrease, whole genome sequencing may eclipse whole exome sequencing, though investigators must also consider the substantial increase in storage space and computational time needed for whole genome sequencing when their sole aim is to assess exonic variants. The higher depths required in somatic studies are likely to favour whole exome sequencing and targeted sequencing in that arena for some time.



Multiplex

Read depth	1x	2x	3x	4x	5x	6x	7x	8x	9x	10x	11x	12x	13x	14x	15x	16x	17x	18x	19x	20x	21x	22x	23x	24x	25x	26x	27x	28x	29x	30x
4 samples per lane	1x	2x	3x	4x	5x	6x	7x	8x	9x	10x	11x	12x	13x	14x	15x	16x	17x	18x	19x	20x	21x	22x	23x	24x	25x	26x	27x	28x	29x	30x
3 samples per lane	1x	3x	4x	5x	7x	8x	9x	11x	1x	13x	15x	16x	17x	19x	20x	21x	23x	24x	25x	27x	28x	29x	31x	32x	33x	35x	36x	37x	39x	40x
2 samples per lane	2x	4x	6x	8x	10x	12x	14x	16x	18x	20x	22x	24x	26x	28x	30x	32x	34x	36x	38x	40x	42x	44x	46x	48x	50x	52x	54x	56x	58x	60x
1 sample per lane	4x	8x	12x	16x	20x	24x	28x	32x	36x	40x	44x	48x	52x	56x	60x	64x	68x	72x	76x	80x	84x	88x	92x	96x	100x	104x	108x	112x	116x	120x

Fig 3 | Top panel: Multiplexing affects sequencing depth. Each matrix represents the sequencing depths across a genomic region for a sample sequenced alongside other samples. Sequential nucleotide bases of a region are shown horizontally; the number of sequence reads assessing each base are shown vertically. Final tallies show the percentages of the targeted region covered at a depth of 20 reads. Only 35% of bases in the target region had ≥20 reads when four samples were multiplexed, whereas 60% and 95% of the bases had ≥20 reads with three and two samples per lane, respectively. Bottom panel: Relation between multiplexing and sequencing coverage of a region. Read depths are shown horizontally and multiplexing vertically. We can estimate the depths likely to be covered at the same percentage for different numbers of multiplexed samples. If a previous, similar experiment with four samples yielded 90% of the target region covered by at least 10 reads, then 20× coverage (over the same 90% of the target region) could be achieved by sequencing two samples per lane.

Example D

Situation—An investigator studying a childhood disease cohort wants to identify high confidence *de novo* mutations (variants in children that are not in either of their parents) using whole exome sequencing in father-mother-child trios.

Application—At the investigator's sequencing center, whole exome sequencing of samples with the same target bait yielded 98% of the exome covered at 16× when multiplexed at four samples per lane. The investigator would like 98% coverage at 20×, so uses fig 3B to find that ~21× coverage over 98% of the exome region will be obtained by sequencing three samples per lane.

If not applied—Using an uninformed multiplex number the investigator might substantially over-sequence the target region, yielding a minimal increase in accuracy for a substantial increase in cost. Even worse, the investigator may under-sequence the target region and not achieve enough depth to detect mutations at many loci.

Adequate sample size for desired power

The sample size planned for any study should have sufficient power to detect a meaningful effect difference (such as a difference in the proportion of people carrying a variant who have or do not have a disease) with statistical significance. Studies with insufficient samples may fail to detect a true association where it exists. Sample size should be determined in advance based on the estimated effect size, the statistical test to be used, and the desired rates of false positives and false negatives. Determining an adequate sample size to assess a hypothesis without wasting financial resources is as crucial to an NGS study as determining the optimal depth.

NGS requires a much lower probability for statistical significance (false positive rate) owing to the total number of simultaneous tests performed. A significance level of 0.05 allows for an average of five in every 100 identified associations to be false. If assessing 30 million exonic bases, this would allow 1.5 million false findings. Multiple testing requires a more stringent P value for significance. The Bonferroni correction (considered the simplest and most stringent multiple comparison correction method) requires a P value ≤0.05 divided by the total number of independent tests performed. For 30 million tests, a standard critical P value of 0.05 equates to a Bonferroni level for significance of $P \leq 1.67 \times 10^{-9}$. Although specific thresholds vary based on non-independence of genetic loci, model assumptions, and use of alternative false discovery rate methods, but a P value of between $1x10^{-7}$ to $1x10^{-8}$ is generally considered necessary for exome-wide significance in NGS studies.[8 22 23]

Because of these heightened significance requirements, investigators should thoughtfully determine the necessary sample size in the initial stages of study design. Clinicians, statisticians, and bioinformaticians should collaborate in planning the study, each bringing a unique skill set to the design process. The sample size calculation for evaluating an estimated effect with an appropriate statistical test, power, and stringent P value can be performed using known formulae, software, or tables (see supplementary web table w1, which contains calculated sample sizes for detecting a difference in proportions with Fisher's exact test[24] in the NGS setting). The proportion of samples with a variant will vary across studies; background mutation rates vary widely across genes, cancers, and study cohorts, reflecting both different intrinsic and extrinsic factors.[25-27] Hence, the proportion of cases and controls having a particular genetic aberration is often not known precisely beforehand, but an informed estimate can help greatly in determining an appropriate sample size.

Example E

Situation—An investigator wants to assess the association between a common disease and the presence of nonsense variants in all genes in an understudied population using whole exome sequencing in both cases and controls.

Application—Based on data from other populations, the investigator wants to detect genes that have a prevalence of nonsense variants in 10% and 50% of unaffected and affected people, respectively. The investigator wants to have 90% power of detecting this effect difference after adjusting for multiple comparisons. For a standard critical P value of 0.05 and a planned Fisher's exact test to be performed on ~20 000 genes, the investigator decides on a conservative P value threshold for significance of 2.5×10^{-6} (0.05/20 000). Using supplementary web table w1, the investigator estimates that a sample size of 90 in each group is necessary.

If not applied—If the investigator sequenced an arbitrary of, for example, 40 cases and 40 controls, the likelihood of identifying such an effect difference would be much lower (the power would be only 20%). In such an underpowered study, a null finding would provide the investigator with little insight into the hypothesis.

Discussion and conclusions

The four study design principles discussed here should be commonplace in designing NGS studies. Other implementations are also important, such as comparison of DNA from the same extraction method and tissue source. Investigators should avoid combining samples of DNA extracted from blood, saliva, and buccal sources as this can result in biased calls of genetic variation, particularly insertions and deletions.[28 29] In general, the more homogeneously that cases and controls are treated throughout the entire sequencing process, the better the experiment.

The genomic assessment possible with NGS is immense. But these studies must conform to basic requirements for good study design to be effective and meaningful. Careful consideration of study objectives and available resources together with increased

attention to study design principles will hopefully improve the rate at which scientific discoveries are made to benefit mankind.

I thank Christopher Ours, William Thomsen, and *The BMJ* reviewers and committee for their helpful suggestions on the draft manuscript.

Contributors and sources: CCM is an assistant professor of pediatrics at the University of Utah, Salt Lake City, USA, where he has taught classes on genomic analysis. He has nearly 10 years' experience in the design, reproducibility assessment, and analysis of genome-wide investigations of common and rare diseases, including cancer. CCM conceived the paper, designed the figures, performed the statistical calculations, and wrote the paper. CCM is the guarantor.

Funding: This manuscript was funded by the pediatric cancer program supported by the Intermountain Healthcare and Primary Children's Hospital Foundations, the University of Utah, Department of Pediatrics, and the Division of Hematology/Oncology (CCM).

Competing interests: CCM has received discounted research products used in genomic experiments from Agilent.

Provenance: Not commissioned; externally peer reviewed.

1 Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;17:333-51. doi:10.1038/nrg.2016.49
2 Head SR, Komori HK, LaMere SA, et al. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* 2014;56:61-4, 66, 68 passim. doi:10.2144/000114133
3 Horak P, Fröhling S, Glimm H. Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls. *ESMO Open* 2016;1:e000094. doi:10.1136/esmoopen-2016-000094
4 Zentner GE, Henikoff S. High-resolution digital profiling of the epigenome. *Nat Rev Genet* 2014;15:814-27. doi:10.1038/nrg3798
5 Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 2014;15:121-32. doi:10.1038/nrg3642
6 Wang Q, Lu Q, Zhao H. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Front Genet* 2015;6:149. doi:10.3389/fgene.2015.00149
7 Auer PL, Reiner AP, Wang G, et al. NHLBI GO Exome Sequencing Project. Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *Am J Hum Genet* 2016;99:791-801. doi:10.1016/j.ajhg.2016.08.012
8 Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014;15:335-46. doi:10.1038/nrg3706
9 McIntyre LM, Lopiano KK, Morse AM, et al. RNA-seq: technical variability and sampling. *BMC Genomics* 2011;12:293. doi:10.1186/1471-2164-12-293
10 Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* 2012;13:705-19. doi:10.1038/nrg3273
11 Torrone DZ, Kuriakose J, Moors K, et al. Reproducibility and intraindividual variation over days in buccal cell DNA methylation of two asthma genes, interferon γ (IFNγ) and inducible nitric oxide synthase (iNOS). *Clin Epigenetics* 2012;4:3. doi:10.1186/1868-7083-4-3
12 Li J, Liu Y, Kim T, Min R, Zhang Z. Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS Comput Biol* 2010;6:e1000910. doi:10.1371/journal.pcbi.1000910
13 Abecasis GR, Altshuler D, Auton A, et al. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061-73. doi:10.1038/nature09534
14 Fu W, O'Connor TD, Jun G, et al. NHLBI Exome Sequencing Project. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 2013;493:216-20. doi:10.1038/nature11690
15 Lek M, Karczewski KJ, Minikel EV, et al. Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285-91. doi:10.1038/nature19057
16 Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. *Applied Linear Statistical Models.* 4th ed. McGraw-Hill Companies, Inc, 1996.
17 Hu Y-J, Liao P, Johnston HR, Allen AS, Satten GA. Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls. *PLoS Genet* 2016;12:e1006040. doi:10.1371/journal.pgen.1006040
18 Shearer AE, Hildebrand MS, Ravi H, et al. Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomic enrichment. *BMC Genomics* 2012;13:618. doi:10.1186/1471-2164-13-618
19 Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 2012;40:e3. doi:10.1093/nar/gkr771
20 Sanders SJ, Murtha MT, Gupta AR, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012;485:237-41. doi:10.1038/nature10945
21 Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 2017;542:433-8. doi:10.1038/nature21062
22 Fadista J, Manning AK, Florez JC, Groop L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet* 2016;24:1202-5. doi:10.1038/ejhg.2015.269
23 Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* 2015;7:16. doi:10.1186/s13073-015-0138-2
24 SAS Institute Inc. *SAS/STAT 13.1 User's Guide.* SAS Institute Inc, 2013.
25 Wu S, Powers S, Zhu W, Hannun YA. Substantial contribution of extrinsic risk factors to cancer development. *Nature* 2016;529:43-7. doi:10.1038/nature16166
26 Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* 2015;347:78-81. doi:10.1126/science.1260825
27 Crompton BD, Stewart C, Taylor-Weiner A, et al. The genomic landscape of pediatric Ewing sarcoma. *Cancer Discov* 2014;4:1326-41. doi:10.1158/2159-8290.CD-13-1037
28 Zhu Q, Hu Q, Shepherd L. The impact of DNA input amount and DNA source on the performance of whole-exome sequencing in cancer epidemiology. *Cancer Epidemiol Biomarkers Prev* 2015;24:1207-13. doi:10.1158/1055-9965.EPI-15-0205
29 Hansen TV, Simonsen MK, Nielsen FC, Hundrup YA. Collection of blood, saliva, and buccal cell samples in a pilot study on the Danish nurse cohort: comparison of the response rate and quality of genomic DNA. *Cancer Epidemiol Biomarkers Prev* 2007;16:2072-6. doi:10.1158/1055-9965.EPI-07-0611

Appendix: Supplementary materials