Mamm Genome (2012) 23:490–498
DOI 10.1007/s00335-012-9402-6

REVIEWS

Next-generation sequencing of experimental mouse strains

Binnaz Yalcin · David J. Adams · Jonathan Flint · Thomas M. Keane

Received: 26 March 2012 / Accepted: 24 May 2012 / Published online: 7 July 2012

Abstract Since the turn of the century the complete genome sequence of just one mouse strain, C57BL/6J, has been available. Knowing the sequence of this strain has enabled large-scale forward genetic screens to be performed, the creation of an almost complete set of embryonic stem (ES) cell lines with targeted alleles for protein-coding genes, and the generation of a rich catalog of mouse genomic variation. However, many experiments that use other common laboratory mouse strains have been hindered by a lack of whole-genome sequence data for these strains. The last 5 years has witnessed a revolution in DNA sequencing technologies. Recently, these technologies have been used to expand the repertoire of fully sequenced mouse genomes. In this article we review the main findings of these studies and discuss how the sequence of mouse genomes is helping pave the way from sequence to phenotype. Finally, we discuss the prospects for using de novo assembly techniques to obtain high-quality assembled genome sequences of these laboratory mouse strains, and what advances in sequencing technologies may be required to achieve this goal.

B. Yalcin
Center for Integrative Genomics, University of Lausanne,
Lausanne, Switzerland
e-mail: Binnaz.Yalcin@unil.ch

B. Yalcin
Institute of Genetics and Molecular and Cellular Biology,
67404 Illkirch, France

D. J. Adams · T. M. Keane (✉)
Wellcome Trust Sanger Institute, Hinxton,
Cambridge CB10 1HH, UK
e-mail: tk2@sanger.ac.uk

D. J. Adams
e-mail: da1@sanger.ac.uk

J. Flint
Wellcome Trust Centre for Human Genetics,
Roosevelt Drive, Oxford OX3 7BN, UK
e-mail: jf@well.ox.ac.uk

Introduction

In recent years, DNA sequencing has undergone a revolution through the development of much higher throughput sequencing technologies resulting in a significant reduction in the cost per base pair (Turner et al. 2009). We have reached the point where it is now possible to sequence the entire genome of a mammalian species for just a tiny fraction of what it cost to generate the raw sequencing data for the mouse reference genome. These second-generation sequencing technologies such as Illumina (Bentley et al. 2008), Roche/454 (Margulies et al. 2005), and SOLiD (Shendure et al. 2005) are based largely on the same principle: sequencing many millions of DNA fragments in parallel (Turner et al. 2009). The sequencing reads produced by these technologies are generally much shorter than capillary sequence reads, a factor that conflates the challenge of analyzing large mammalian genomes (Pop and Salzberg 2008).

We used second-generation sequencing technologies to deeply sequence 17 mouse strains on the Illumina platform (Keane et al. 2011; Yalcin et al. 2011). In this review we describe the different types of sequence variation uncovered, with specific emphasis on structural variation, and discuss the implications of our findings for understanding how sequence variation influences phenotypic differences. Finally, we examine the prospects for using second- or third-generation sequencing technologies to create improved high-quality (Chain et al. 2009) genome sequences for these mouse strains.

Identification of SNPs and short indels

The raw sequence for our study of the 17 mouse strains was generated on the Illumina GAII platform (Bentley et al. 2008), with reads of between 54 and 108 bp generated from both ends of DNA fragments of 300–500 bp in size. When these reads were aligned to the reference strain (C57BL/6J; MGSCv37 assembly), 13–23 % of the reference genome assembly could not be confidently accessed due to the presence of highly divergent sequence or high copy-repeated sequences that were longer than the sequence reads and fragment size (such as transposable elements, telomeric repeats, centromeres, or low-complexity regions) (Flicek and Birney 2009).

In the mouse genome, and indeed in other vertebrate genomes, the simplest and most prevalent type of molecular variation is the single nucleotide polymorphism (SNP). The algorithms for calling SNPs scan across the reference genome observing the aligned read bases at each position, and then use read depth and base quality to identify sequence mismatches with high accuracy (Pop and Salzberg 2008). Our analysis found a total of 56.7 M SNP sites, but the number of SNPs varied considerably among strains, ranging from just a few thousand in the C57BL/6NJ strain to 35.4 M in SPRET/EiJ. The major denominator for the number of SNPs discovered was the genetic distance of the mouse strain from the reference C57BL/6J genome. A combination of three SNP calling algorithms were used (SAMtools (Li et al. 2009), GATK (McKenna et al. 2010), and QCALL (Le and Durbin 2011)), with the final set of SNPs consisting of sites that were identified by at least two of the callers. In agreement with findings from the human 1000 Genomes pilot project where a majority voting scheme was employed to merge SNP genotypes (1000 Genomes Project Consortium 2010), this strategy was found to minimize the false discovery rate while maintaining high sensitivity. Small insertions and deletions (indels) of 1–100 bp were also detected using a combination of Dindel (Albers et al. 2011) and also by carrying out de novo assembly of the reads and comparing the resulting contigs to the reference genome assembly (Keane et al. 2011). Overall there were approximately six times fewer indels than SNPs, and it was found that the indel calls were of lower sensitivity and specificity than SNP calls owing to the complexity of calling these variants from short read sequences.

The accuracy of SNP and indel calls was established by comparing variant calls to 16.3 Mbp of finished BAC sequences from the NOD/ShiLtJ strain. The NOD/ShiLtJ BAC sequence represented a unique resource of high-quality finished sequence that allowed us to robustly assess our false-negative and false-positive rates. In inaccessible regions, the 13–23 % of the reference genome where we were unable to unequivocally place sequence reads, we found a threefold enrichment for sequence variants, implying that current sequencing technologies miss at least 30 % of sequence variation. However, it remains unclear how much of this missing variation is functional as the inaccessible regions of the genome are replete with low complexity, simple repeats, and high copy repetitive elements such as long interspersed nuclear elements (LINEs). The number of SNP variants we discovered from sequencing the 17 mouse genomes represented a sevenfold increase in the number of these variants in public databases such as dbSNP.

Interestingly, a significant subset of the SNP (0.12 M) and indel (0.005 M) positions discovered in our analysis resulted in amino acid substitutions, highlighting the diversity in coding sequence between mouse strains.

Identification of structural variation

We defined structural variation as sequence variants that are greater than 100 bp. Structural variants (SVs) are an important source of sequence variation, in both human (Conrad et al. 2010; Feuk et al. 2006; Kidd et al. 2008; Mills et al. 2011; Redon et al. 2006), and mouse (Yalcin et al. 2011; Agam et al. 2010; Akagi et al. 2008; Cahan et al. 2009; Cutler et al. 2007; Graubert et al. 2007; Henrichsen et al. 2009; Quinlan et al. 2010) genomes. They include insertions, retrotransposon elements, inversions, segmental duplications, and other genomic rearrangements.

The extent of structural variation in the mouse genome was first demonstrated using differential hybridization of genomic DNA to oligonucleotide arrays [array comparative genome hybridization (CGH)] (Cahan et al. 2009; Cutler et al. 2007; Graubert et al. 2007; Henrichsen et al. 2009). While array CGH methods can interrogate hundreds of genomes, they are blind to some SV types such as those that are copy-number neutral and rarely provide breakpoint resolution. Furthermore, array CGH is generally unable to detect SVs smaller than 5 kbp and are poorly reproducible between studies (Agam et al. 2010). Previous array CGH analyses estimated that the proportion of the mouse genome affected by SVs ranged from 3 % (Cahan et al. 2009) to over 10 % (Henrichsen et al. 2009).

The greater sensitivity and specificity of next-generation sequencing technologies represent a significant advance on array-based methods of SV identification. Our recent catalog of structural variation contains far more SVs than previously identified (Nellaker et al. 2012; Yalcin et al. 2011). We found structural variants at 281,243 unique sites, amounting to 711,920 SVs in 13 classical and 4 wild-derived inbred strains of mice, affecting 1.2 and 3.7 % of the genome, respectively. The majority of SVs are less than 1 kbp in size, below the level amenable to detection by array CGH methods.

Methods to localize SVs using next-generation sequencing are based on paired-end mapping (PEM) (Medvedev et al. 2009) (also reviewed in (Alkan et al. 2011)): two short paired-reads from both extremities of a segment of DNA (the insert) and at an approximately known distance are sequenced and mapped back to the reference genome. Typically, variation in the expected number of reads mapping to the reference sequence is used to identify copy number variation, while deviations from the expected distance between reads, and the orientation of reads, are used to determine the type of structural variant such as deletions and inversions.

In the past few years, a plethora of software tools have been developed to detect SVs from next-generation sequencing data. These tools exploit (1) read-pair, (2) split-read, (3) read depth, and (4) sequence assembly information. A summary of software tools is provided for each of these methods in Table 1. Algorithms that exploit read-pair (1000 Genomes Project Consortium 2010; Mills et al. 2011; Quinlan et al. 2010; Keane, RetroSeq; Chen et al. 2009; Hormozdiari et al. 2009; Hormozdiari et al. 2010; Hormozdiari et al. 2011; Korbel et al. 2009; Lee et al. 2009; Qi and Zhao 2011; Zeitouni et al. 2010) can detect four types of SVs (deletions, insertions, inversions, and tandem duplications). They look for read-pairs that are anomalously aligned to the reference genome, e.g., reads that are either too far apart or in the wrong orientation. When one of the paired-reads is mapped to the reference genome and the other is unmapped, this suggests a large insertion.

In the split-read approach (Albers et al. 2011; Emde et al. 2012; Karakoc et al. 2011; Ye et al. 2009; Zhang et al. 2011), one of the paired-reads is mapped to the reference genome, acting as an anchor, while the other encompasses the structural variant, typically a small insertion. Additionally, the high coverage of next-generation sequencing makes it possible to detect copy number changes using the read-depth approach (Abyzov et al. 2011; Klambauer et al. 2012; Medvedev et al. 2010; Simpson et al. 2010; Yoon et al. 2009). Assembly algorithms (Mills et al. 2011; Chaisson et al. 2009; Gnerre et al. 2011; Hajirasouliha et al. 2010; Li et al. 2010; Simpson et al. 2009) have the most power to detect SVs at base-pair resolution; however, they also miss considerable variation, especially at complex genomic regions (Alkan et al. 2011).

None of the algorithms is ideal, and none deals well with complex SVs that consist of a combination of rearrangements (such as an insertion abutting a deletion or an inversion within a gain) (reviewed in (Quinlan and Hall 2012)). In a recent study (Yalcin et al. 2012), we manually examined the whole of chromosome 19 for structural variation using the short read visualization tool LookSeq (Manske and Kwiatkowski 2009). We found greater diversity and complexity in SVs than had previously been reported. The manually curated set of SVs provided a benchmark for developing a method to call complex SVs at a genome-wide level (SVMerge (Wong et al. 2010)). It should be noted that SVMerge is the first tool, to date, that can effectively call complex SVs (Table 1). To study the full spectrum of SVs, future algorithms need to consider the complex forms of PEM patterns, described in (Yalcin et al. 2012).

Functional impact of structural variation

Although twice more base pairs are affected as a consequence of structural variation than single nucleotide mutation, it remains unclear to what extent structural variants contribute to quantitative phenotypic differences. On the one hand, there have been some reports that common SVs are less likely than common SNPs to contribute to phenotypic variation (Keane et al. 2011; Conrad et al. 2010). On the other hand, several studies have provided remarkable estimates of the contribution of SVs to variation in transcript abundance: estimates ranged from 10 to 74 % (Yalcin et al. 2011; Cahan et al. 2009; Henrichsen et al. 2009; Stranger et al. 2007). It has also been reported that structural variants can influence gene expression up to 500 kbp from their margins (Henrichsen et al. 2009). Since gene expression variation is believed to contribute to variation in phenotypes at a whole-organism level (Schadt et al. 2005), results from these studies might indicate that the phenotypic impact of SVs is large.

Our genome-wide catalog of SVs was used in two ways to address the extent to which SVs affect phenotypic differences. The first used results from genome-wide association studies in an outbred population of mice, the Northport heterogeneous stock (HS) mice (Demarest et al. 2001; Talbot et al. 1999). The Northport HS mice are animals derived from eight of the sequenced strains (A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J, and LP/J) (Keane et al. 2011). Because many recombinants have accumulated since the creation of the HS population, mapping resolution of quantitative trait loci (QTLs) is high (to an average region of 3 Mbp). The HS population is not only unique for its high mapping resolution but also for the large number of QTLs that have already been mapped for a diversity of traits (about 100 traits) (Valdar et al. 2006). Since the HS mice are derived from eight fully sequenced strains, they can be used to assess the impact of genomic variants such as SVs on phenotypic differences (Yalcin and Flint 2012).

To do this, we applied a test of functionality (Yalcin et al. 2005) that allowed us to discriminate between SVs that are likely to be functional and those that are not. We found that the larger the effect, the more likely it is to arise from a structural variant (Keane et al. 2011). However,

Table 1 A summary of software tools to detect simple and complex SVs

Method	Software	Detectable SV types					Reference
		Del	Ins	Inv	Dup	Complex	
Read-pair	BreakDancer	✔	✔	✔	✔		(Chen et al. 2009)
	HYDRA	✔	✔	✔	✔		(Quinlan et al. 2010)
	inGAP-sv	✔	✔	✔	✔		(Qi and Zhao 2011)
	MoDIL	✔	✔	✔	✔		(Lee et al. 2009)
	PEMer	✔	✔	✔	✔		(Korbel et al. 2009)
	RetroSeq		✔				(Keane, RetroSeq)
	SPANNER	✔	✔	✔	✔		(1000 Genomes Project Consortium 2010; Mills et al. 2011)
	SVDetect	✔	✔	✔	✔		(Zeitouni et al. 2010)
	VariationHunter	✔	✔	✔	✔		(Hormozdiari et al. 2009; Hormozdiari et al. 2010; Hormozdiari et al. 2011)
Split-read	Dindel	✔	✔				(Albers et al. 2011)
	Pindel	✔	✔				(Ye et al. 2009)
	SplazerS	✔	✔				(Emde et al. 2012)
	Splitread	✔	✔				(Karakoc et al. 2011)
	SRiC	✔	✔				(Zhang et al. 2011)
Read depth	cnD	✔			✔		(Simpson et al. 2010)
	cn.MOPS	✔			✔		(Klambauer et al. 2012)
	CNVer	✔			✔		(Medvedev et al. 2010)
	CNVnator	✔			✔		(Abyzov et al. 2011)
	EWT	✔			✔		(Yoon et al. 2009)
Assembly	ABySS	✔	✔	✔	✔		(Simpson et al. 2009)
	ALLPATHS-LG	✔	✔	✔	✔		(Gnerre et al. 2011)
	EULER-USR	✔	✔	✔	✔		(Chaisson et al. 2009)
	NovelSeq	✔	✔	✔	✔		(Hajirasouliha et al. 2010)
	SOAPdenovo	✔	✔	✔	✔		(Li et al. 2010)
	TIGRA	✔	✔	✔	✔		(Mills et al. 2011)
Meta caller	SVMerge	✔	✔	✔	✔	✔	(Wong et al. 2010)

there are very few QTLs that are likely to be due to a structural variant: we identified just 12 QTLs where a structural variant overlapped a gene and where the effect size was in the top 5 % of the distribution. Table 2A lists these genes and the putative phenotypes with which they are associated. In one case, we used complementation (Mackay 2004) of a deletion of the *H2–Ea* promoter to confirm the effect of this SV on a T cell phenotype (Yalcin et al. 2010). In another case, we had evidence in favor of a causative role for an insertion in the promoter of one gene (*Eps15*) that abolished gene expression. We found that *Eps15* knockout mice exhibited a significantly lower locomotor activity compared to matched wild-type mice, indicating that the insertion is likely the cause of the QTL.

The second way in which a genome-wide catalog of SVs can be used to assess the functional impact of SVs is by identifying variants that remove a coding segment of a gene, effectively creating a null or altered allele. Again these are relatively few. A summary of genes containing SVs affecting coding regions is provided in Table 2B. Most of these SVs have been newly identified (Yalcin et al. 2011), and a small number were already known (Best et al. 1996; Boyden et al. 2008; Nelson et al. 2005; Persson et al. 1999; Wu et al. 2010). These SVs, with large effects on a phenotype, are the equivalent of rare variants found in human populations. In the mouse, these SVs are rare relative to their abundance in the genome; however, they provide, for the first time, biological insights into the influence of these events on phenotype.

Complex molecular architecture of SVs

Because of their high breakpoint accuracy, our genome-wide catalog of SVs not only expands knowledge of the molecular architecture of SVs, it also allows inferring a SV mechanism of formation with a high degree of precision. We know that more than half of the SVs are caused by retrotransposition of

Table 2 Mouse genes affected by a structural variant correlated to a phenotype

Gene	SV event	Chr	SV start	SV stop	Phenotype	Reference
A. Genes with SV associated with quantitative traits						
4921524J17Rik	LINE Ins	8	87957244	87957245	Red cells: mean cellular volume	(Yalcin et al. 2011)
Eps15	IAP Ins	4	108951263	108951264	Home cage activity	(Yalcin et al. 2011)
Fcer1a	Ins	1	175158884	175158885	Mean platelet volume	(Yalcin et al. 2011)
Gm6320	Del	13	113783196	113783359	Hippocampus cellular proliferation	(Yalcin et al. 2011)
Grin3a	Del	4	49690362	49690363	Hippocampus cellular proliferation	(Yalcin et al. 2011)
H2–Ea	Del	17	34483681	34483682	T-cells: CD4/CD8 ratio	(Yalcin et al. 2011; Yalcin et al. 2010)
Nnt	Del	13	120164268	120164269	Glucose intolerance	(Freeman et al. 2006)
Sec23b	SINE Ins	2	144402760	144402971	OFT total activity	(Yalcin et al. 2011)
Snrnp40	SINE Ins	4	130038388	130038389	T-cells: % CD3	(Yalcin et al. 2011)
Tmc3	IAP Ins	7	90731819	90731820	Wound healing	(Yalcin et al. 2011)
Tmem104	Del	11	115106127	115106250	Serum urea concentration	(Yalcin et al. 2011)
Trim30b	Del	7	111504989	111505193	Red cells: mean cellular hemoglobin	(Yalcin et al. 2011)
Trim5	Ins	7	111397607	111479433	Red cells: mean cellular hemoglobin	(Yalcin et al. 2011)
B. Genes with SV affecting coding regions						
Amd2	Ins	18	64607747	64609669	Biosynthesis of polyamines	(Yalcin et al. 2011; Persson et al. 1999)
Defb8	Ins + Del	8	19447465	19450575	Infection and immunity	(Yalcin et al. 2011; Bauer et al. 2001)
Fam110c	VNTR	12	31759321	31759461	Cell migration	(Yalcin et al. 2011)
Fcrl5	Del	3	87245084	87245947	Infection and immunity	(Yalcin et al. 2011)
Fv1	Del	4	147244398	147245739	Infection and immunity	(Yalcin et al. 2011; Best et al. 1996)
Klrb1a	Del	6	128559593	128559740	Infection and immunity	(Yalcin et al. 2011)
Klri2	Del	6	129689526	129691211	Infection and immunity	(Yalcin et al. 2011)
Krtap16-1	VNTR	16	88874294	88874392	Hair formation	(Yalcin et al. 2011)
Krtap5-5	VNTR	7	149415121	149415210	Hair formation	(Yalcin et al. 2011)
Nes	VNTR	3	87780530	87780662	Brain development	(Yalcin et al. 2011)
Nlrp1c	Ins	11	71046193	71101410	Embryonic development	(Yalcin et al. 2011)
Olfr1055	IAP Ins	2	86179898	86186982	Olfactory	(Yalcin et al. 2011)
Olfr234	Del	15	98328544	98328861	Olfactory	(Yalcin et al. 2011)
Olfr913	Del	9	38402589	38403498	Olfactory	(Yalcin et al. 2011)
Pglyrp3	Del	3	91831862	91835385	Infection and immunity	(Yalcin et al. 2011)
Rtp3	VNTR	9	110889280	110889465	Bone density	(Yalcin et al. 2011)
Skint4,3,9	Ins	4	111731004	112272814	Infection and immunity	(Yalcin et al. 2011; Boyden et al. 2008)
Soat1	Del	1	158394620	158401436	Hair interior defects	(Yalcin et al. 2011; Wu et al. 2010)
Tas2r103	Del	6	132985563	132986696	Taste	(Yalcin et al. 2011; Nelson et al. 2005)
Tas2r120	Del + Ins	6	132580541	132613777	Taste	(Yalcin et al. 2011; Nelson et al. 2005)
Trim5,12a	Ins	7	111397607	111479433	Infection and immunity	(Yalcin et al. 2011; Tareen et al. 2009)
Ugt2b38	Del	5	87850554	87854999	Metabolism	(Yalcin et al. 2011)
Zfp607	Del	7	28646761	28671650	DNA-binding	(Yalcin et al. 2011)
Zfp872	VNTR	9	22004856	22005023	DNA-binding	(Yalcin et al. 2011)

Box 1

Resource name	Description	URL
Mouse genomes project	Wellcome Trust Sanger Institute mouse genomes project webpage with details of the mouse strains sequenced and how to get the data	http://www.sanger.ac.uk/resources/mouse/genomes/
Mouse genomes project browser	The website for querying and downloading lists of SNPs, indels, and structural variants	http://www.sanger.ac.uk/cgi-bin/modelorgs/mousegenomes/snps.pl
dbSNP	All SNP and indel variants from the 17 strains have been submitted to dbSNP under handle "SC_MOUSE_GENOMES"	http://www.ncbi.nlm.nih.gov/projects/SNP/
DGVa	Database of genomic variants archive (DGVa) is a repository that provides archiving, accessioning, and distribution of publicly available genomic structural variants. All structural variants from the 17 strains have been submitted under accession estd118 and estd185	http://www.ebi.ac.uk/dgva/

LINEs (25 % of SVs), SINEs (short interspersed nuclear elements; 15 %), and LTRs (long terminal repeats; 14 %), followed by VNTRs (variable-number tandem repeats; 15 %), and pseudogenes (2 %) (Yalcin et al. 2011).

By characterizing the sequence features around SV breakpoints, we found that about a quarter of SVs have smaller rearrangements at their breakpoints, such as a microinsertion or a microdeletion at the breakpoint of a larger variant (Yalcin et al. 2012). For example, two alleles have been reported for a β-defensin gene (*Defb8*) that differ by 3 bp changes in the second exon (Bauer et al. 2001; Taylor et al. 2009). We found that, in fact, these documented exonic changes are linked to a previously undetected 3,192 bp deletion (Yalcin et al. 2011).

It is acceptable to assume that the complex molecular architecture of SV (microstructures at SV breakpoints) will correlate with complex mechanisms of SV formation. Two mechanisms, a DNA replication fork stalling and template switching (FoSTeS) and a microhomology-mediated break-induced replication (MMBIR), have been proposed to generate such complex SVs in the human genome (Zhang et al. 2009). It could be that the complex SVs we see in the mouse genome (about 25 % of SVs) have also formed through mutational forces during DNA replication.

However, as highlighted previously, there are real limitations in the methods of SV detection with complex molecular architecture. Ideally, sequencing of larger fragments of DNA or, even better, complete de novo assembly of the genome would typically be required to resolve the full spectrum of complex architecture of structural variants.

Prospects for full-genome sequences

While our high-resolution catalogs of sequence variation advanced studies correlating genotype to phenotype, the ultimate goal is to obtain the complete genomic sequence of all common laboratory mouse strains. As a first step toward this goal, the sequencing reads generated by the Mouse Genomes Project have been put through de novo assembly using the Velvet assembler (Zerbino and Birney 2008) and preliminary draft genomes are available for download from the project FTP site (see Box 1). Preliminary analysis of these draft assemblies shows that approximately 90 % of the coding regions of the strains can be found in the assemblies of the strains, although this number is lower for the wild-derived strains. Clearly much work remains to bring these assemblies up to a standard approaching that of the C57BL/6J reference genome.

Future work

The Mouse Genomes Project produced an unprecedented amount of raw sequencing data across 17 mouse strains. The analysis of these data has painted the most comprehensive picture of molecular variation in the mouse genome to date. However, due to limitations in second-generation sequencing technologies, up to 30 % more sequence variation remains to be discovered. To this end, efforts are underway to resequence the strains with longer and higher-quality reads from newer versions of second-generation sequencing technologies. However, the key to discovering the complete set of sequence variants will be the development of third-generation sequencing technologies capable of producing much longer read sequences (multiple kbp in size) so that we can interrogate parts of the genome that are out of reach for current technologies (Schadt et al. 2010).

The ultimate goal of the de novo assembly efforts is to produce full-genome sequences of the 17 strains of quality comparable to that of the reference genome. It has been shown that to go from a set of assembled contigs to larger scaffolds of hundreds of kilobases, sequencing from the ends of large fragments of varying sizes is required (Gnerre et al. 2011). Long-fragment sequencing remains challenging with second-generation sequencing technologies; primarily due to difficulties in producing sufficiently diverse sequencing libraries and reliable methods are still under development (Van Nieuwerburgh et al. 2012). De novo assembly is an area that will benefit greatly from the development of third-generation sequencing technologies capable of producing much longer read lengths.

The C57BL/6J mouse has been the mouse reference genome since the turn of century. However, as we produce improved de novo assemblies from the newly sequenced strains, we can use the novel sequence haplotypes found in subsets of the 17 strains and not found in the reference genome to define the mouse pan-genome reference (Dunn et al. 2012; Li et al. 2010; Muzzi and Donati 2011). The goal of creating this pan-genome would be to reduce the reference bias that affects many experiments and allow for the discovery of sequence variation shared among subsets of strains and not found in C57BL/6J.

Acknowledgments This work was supported by the Medical Research Council, UK, and the Wellcome Trust. Binnaz Yalcin is supported by an EMBO Fellowship.

References

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21:974–984

Agam A, Yalcin B, Bhomra A, Cubin M, Webber C, Holmes C, Flint J, Mott R (2010) Elusive copy number variation in the mouse genome. PLoS One 5:e12839

Akagi K, Li J, Stephens RM, Volfovsky N, Symer DE (2008) Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. Genome Res 18:869–880

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R (2011) Dindel: accurate indel calls from short-read data. Genome Res 21:961–973

Alkan C, Coe BP, Eichler EE (2011a) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363–376

Alkan C, Sajjadian S, Eichler EE (2011b) Limitations of next-generation genome sequence assembly. Nat Methods 8:61–65

Bauer F, Schweimer K, Kluver E, Conejo-Garcia JR, Forssmann WG, Rosch P, Adermann K, Sticht H (2001) Structure determination of human and murine beta-defensins reveals structural conservation in the absence of significant sequence similarity. Protein Sci 10:2470–2479

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Best S, Le Tissier P, Towers G, Stoye JP (1996) Positional cloning of the mouse retrovirus restriction gene Fv1. Nature 382:826–829

Boyden LM, Lewis JM, Barbee SD, Bas A, Girardi M, Hayday AC, Tigelaar RE, Lifton RP (2008) Skint1, the prototype of a newly identified immunoglobulin superfamily gene cluster, positively selects epidermal gammadelta T cells. Nat Genet 40:656–662

Cahan P, Li Y, Izumi M, Graubert TA (2009) The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. Nat Genet 41:430–437

Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C et al (2009) Genomics. Genome project standards in a new era of sequencing. Science 326:236–237

Chaisson MJ, Brinza D, Pevzner PA (2009) De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res 19:336–346

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP et al (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6:677–681

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P et al (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712

Cutler G, Marshall LA, Chin N, Baribault H, Kassner PD (2007) Significant gene content variation characterizes the genomes of inbred mouse strains. Genome Res 17:1743–1754

Demarest K, Koyner J, McCaughran J Jr, Cipp L, Hitzemann R (2001) Further characterization and high-resolution mapping of quantitative trait loci for ethanol-induced locomotor activity. Behav Genet 31:79–91

Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012) Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. Genome Res 22(5):908–924

Emde AK, Schulz MH, Weese D, Sun R, Vingron M, Kalscheuer VM, Haas SA, Reinert K (2012) Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. Bioinformatics 28(5):619–627

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods 6:S6–S12

Freeman HC, Hugill A, Dear NT, Ashcroft FM, Cox RD (2006) Deletion of nicotinamide nucleotide transhydrogenase: a new quantitative trait locus accounting for glucose intolerance in C57BL/6 J mice. Diabetes 55:2153–2156

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108:1513–1518

Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM, Ley TJ (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet 3:e3

Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC (2010) Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 26:1277–1283

Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A (2009) Segmental copy number variation shapes tissue transcriptomes. Nat Genet 41:424–429

Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res 19:1270–1278

Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC (2010) Next-generation Variation-Hunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 26:i350–i357

Hormozdiari F, Hajirasouliha I, McPherson A, Eichler EE, Sahinalp SC (2011) Simultaneous structural variation discovery among multiple paired-end sequenced genomes. Genome Res 21:2203–2212

Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, Mark K, Rieder MJ, Nickerson DA, Eichler EE (2011) Detection of structural variants and indels within exome data. Nat Methods 9:176–178

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M et al (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289–294

Keane T, (2012) RetroSeq: A tool for discovery and genotyping of transposable elements from short read alignments. Available at https://githubcom/tk2/RetroSeq. Accessed 1 July 2012

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F et al (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res 40(9):e69

Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10:R23

Le SQ, Durbin R (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res 21:952–960

Lee S, Hormozdiari F, Alkan C, Brudno M (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat Methods 6:473–474

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAM tools. Bioinformatics 25:2078–2079

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al (2010a) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J et al (2010b) Building the sequence map of the human pan-genome. Nat Biotechnol 28:57–63

Mackay TF (2004) Complementing complexity. Nat Genet 36: 1145–1147

Manske HM, Kwiatkowski DP (2009) LookSeq: a browser-based viewer for deep sequencing data. Genome Res 19:2125–2132

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6:S13–S20

Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M (2010) Detecting copy number variation with mated short reads. Genome Res 20:1613–1622

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK et al (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65

Muzzi A, Donati C (2011) Population genetics and evolution of the pan-genome of *Streptococcus pneumoniae*. Int J Med Microbiol 301:619–622

Nellaker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, Ponting CP (2012) The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol 13:R45. doi:10.1186/gb-2012-13-6-r45

Nelson TM, Munger SD, Boughter JD Jr (2005) Haplotypes at the *Tas2r* locus on distal chromosome 6 vary with quinine taste sensitivity in inbred mice. BMC Genet 6:32

Persson K, Heby O, Berger FG (1999) The functional intronless S-adenosylmethionine decarboxylase gene of the mouse (Amd-2) is linked to the ornithine decarboxylase gene (*Odc*) on chromosome 12 and is present in distantly related species of the genus Mus. Mamm Genome 10:784–788

Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24:142–149

Qi J, Zhao F (2011) inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. Nucleic Acids Res 39:W567–W575

Quinlan AR, Hall IM (2012) Characterizing complex structural variation in germline and somatic genomes. Trends Genet 28(1):43–53

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res 20:623–635

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W et al (2006) Global variation in copy number in the human genome. Nature 444:444–454

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C et al (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19:R227–R240

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123

Simpson JT, McIntyre RE, Adams DJ, Durbin R (2010) Copy number variant detection in inbred strains from short read sequence data. Bioinformatics 26:565–567

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C et al (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853

Talbot CJ, Nicod A, Cherny SS, Fulker DW, Collins AC, Flint J (1999) High-resolution mapping of quantitative trait loci in outbred mice. Nat Genet 21:305–308

Tareen SU, Sawyer SL, Malik HS, Emerman M (2009) An expanded clade of rodent Trim5 genes. Virology 385:473–483

Taylor K, Rolfe M, Reynolds N, Kilanowski F, Pathania U, Clarke D, Yang D, Oppenheim J, Samuel K, Howie S et al (2009)

Defensin-related peptide 1 (Defr1) is allelic to Defb8 and chemoattracts immature DC and CD4 + T cells independently of CCR6. Eur J Immunol 39:1353–1360

Turner DJ, Keane TM, Sudbery I, Adams DJ (2009) Next-generation sequencing of vertebrate experimental organisms. Mamm Genome 20:327–338

Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet 38:879–887

Van Nieuwerburgh F, Thompson RC, Ledesma J, Deforce D, Gaasterland T, Ordoukhanian P, Head SR (2012) Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. Nucleic Acids Res 40:e24

Wong K, Keane TM, Stalker J, Adams DJ (2010) Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol 11:R128

Wu B, Potter CS, Silva KA, Liang Y, Reinholdt LG, Alley LM, Rowe LB, Roopenian DC, Awgulewitsch A, Sundberg JP (2010) Mutations in sterol O-acyltransferase 1 (Soat1) result in hair interior defects in AKR/J mice. J Invest Dermatol 130:2666–2668

Yalcin B, Flint J (2012) Association studies in outbred mice in a new era of full genome sequencing. Mamm Genome (to be appear)

Yalcin B, Flint J, Mott R (2005) Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. Genetics 171:673–681

Yalcin B, Nicod J, Bhomra A, Davidson S, Cleak J, Farinelli L, Osteras M, Whitley A, Yuan W, Gan X et al (2010) Commercially available outbred mice for genome-wide association studies. PLoS Genet 6(9):e1001085

Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellaker C, Goodstadt L, Nicod J, Bhomra A et al (2011) Sequence-based characterization of structural variation in the mouse genome. Nature 477:326–329

Yalcin B, Wong K, Bhomra A, Goodson M, Keane T, Adams DJ, Flint J (2012) The fine-scale architecture of structural variants in 17 mouse genomes. Genome Biol 13(3):R18

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19:1586–1592

Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoixne P, Nicolas A, Delattre O, Barillot E (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 26:1895–1896

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR (2009) The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 41:849–853

Zhang ZD, Du J, Lam H, Abyzov A, Urban AE, Snyder M, Gerstein M (2011) Identification of genomic indels and structural variations using split reads. BMC Genomics 12:375